Trípodos, número 39 | Barcelona 2016 | 53-69
ISSN: 1138-3305

Rebut / Received: 10/09/16
Acceptat / Accepted: 24/11/16

Unmasking Trolls: Political Discussion on Twitter during the Parliamentary Election in Catalonia

Desemmascarant els trols: Discussió política a Twitter durant les eleccions al Parlament de Catalunya

Eulàlia Puig Abril
University of Illinois at Chicago (United States)

In recent years, Twitter has become a popular online platform where citizens can discuss politics. However, these conversations may take an uncivil turn, and the consequences could be damaging to democracy. One such uncivil behavior on Twitter is trolling, a disruptive online activity geared toward luring others into pointless and time-consuming discussion. Disruptive messages can have severe consequences for the deliberative system in democratic societies and could frustrate the development of a public sphere. This study explores Twitter trolls during the contentious 2012 Parliamentary election in Catalonia, an autonomous region in Spain. Using Gnip's firehose, all tweets containing the word "independence" or the hashtag "#25N" from the four weeks preceding the election were captured for analysis, which generated a corpus of 325,888 tweets. Results based on automated and manual content analysis show that the prevalence of successful trolls was at 0.01%, thus indicating that they were scant despite the contro-

En els últims anys, Twitter s'ha convertit en una plataforma popular en la qual els ciutadans poden parlar de política. No obstant això, les converses a Twitter poden tornar-se incivils, amb conseqüències que podrien ser perjudicials per a la democràcia. Una d'aquestes conductes incivils a Twitter es l'anomenat troling, una activitat disruptiva dirigida a atraure altres en una discussió llarga i sense sentit en el marc d'internet. Aquests missatges pertorbadors poden tenir greus conseqüències per al sistema deliberatiu en les societats democràtiques i podrien frustrar el desenvolupament d'una esfera pública. Aquest estudi explora els trols de Twitter durant les polèmiques eleccions al Parlament de Catalunya el 2012. A través de Gnip, totes les piulades que contenen la paraula "independència" o el hashtag "# 25N" durant les quatre setmanes anteriors a les eleccions van ser capturades per a l'anàlisi, la qual cosa va generar un cos de 325.888 piulades. Els resultats basats en anàlisi de contingut automàtic i manual mostren

versy surrounding the issue of independence. The study also contributes to the identification of successful Twitter trolls and discusses the democratic implications of trolling on Twitter.

Key words: *Twitter, trolls, political discussion, elections, Catalonia.*

que la presència de trols exitosos va ser del 0,01%, cosa que indica que els trols van ésser escassos tot i la controvèrsia de la qüestió independentista. L'estudi també contribueix a la identificació de trols exitosos a Twitter i discuteix les implicacions democràtiques del troling a Twitter.

Paraules clau: *Twitter, trols, discussió política, eleccions, Catalunya.*

Twitter's potential as a tool for political discussion has been validated in several countries (Conover *et al.*, 2011; Larsson and Moe, 2011; Hosch-Dayican *et al.*, 2014) including Spain (Guerrero-Solé, Corominas-Murtra and López-González, 2014). Still, like other types of online discussion, Twitter's democratic potential could be overshadowed by the lack of face-to-face communication, the opportunity to create multiple accounts, and anonymity, which has been shown to lead to incivility (Papacharissi, 2004; Santana, 2013).

One particular incivity is Twitter trolling (Delclós Juanola, 2013; Mooney, 2014). Trolling is intentionally disruptive online behavior embedded in an online discussion. Trolling's end result is wasting everyone's time, but without making this purpose obvious —unlike flaming (Hmielowski, Hutchens and Cicchirillo, 2014)—. Twitter trolling may halt an entire discussion or prevent future ones, especially with controversial issues. As such, trolling is considered uncivil, disruptive, and with negative consequences for online discussion (Golder and Donath, 2004).

Because corralling a Twitter discussion for analysis is challenging (Bruns, 2012), we know little about the extent of trolling on Twitter, its successes, or who the trolls are. This study explores the existence and behaviors of *successful* Twitter trolls during the electoral campaign for the 2012 Parliamentary election in Catalonia, an autonomous region in Spain. This campaign had parties centered on the issue of Catalonia's quest for independence from Spain (Guerrero-Solé, Corominas-Murtra and López-González, 2014), making the election particularly controversial. During that time, Twitter was gaining prominence as a "forum" for political talk (Doval Avendaño and Martínez Rodríguez, 2012), but there was a feared anticipation that individuals on either side of the independence debate —in Catalonia or outside of it— would fuel incivility on Twitter featuring trolls (Pont and Capdevila, 2012).

The inquiry centers on automated and human content analysis (Abril, Szczypka and Emery, 2017)of all relevant tweets (N = 325,888) collected during the

four weeks before Election Day via Gnip's Twitter firehose (Bruns, 2012). The advantage of using Gnip's firehose and not the API (Twitter's Application Program Interface) is that the firehose captures all the tweets sent by all public users (e.g., Emery *et al*., 2014) while the API only distributes an unpredictable fraction of tweets (<https://dev.twitter.com/docs/streaming-apis>), making it an unreliable source (Kim, Huang and Emery, 2016).

Considering the importance of political talk for democracy (Huckfeldt, Johnson, and Sprague, 2004; Mansbridge, 1999), even when there is disagreement (Abril and Rojas, 2015), and potential threats (Yan *et al*., 2016), ever on Twitter (Geiger, 2016), I argue that the analysis of Twitter trolling activity, its patterns, and its success may lead to a better understanding of the democratic constraints of online discussion in microblogging sites. Moreover, since this may be one of the first attempts at exploring trolling on Twitter for a contentious issue, this study advances a new route to identifying successful trolling activity. Study results may also reveal how Twitter users resolve the global discussion of a local issue in real time —just before an election— and in a language other than English, an aspect of trolling and incivility that is warranted (Phillips, 2015).

ONLINE INCIVILITY AND TROLLING

Civility is thought to play a central role in political discussion among democratic societies. Although sometimes civility is confused with politeness (for a review, see Papacharissi, 2004), civility implies consideration for the consequences of one's behavior; that is, "…respect for the collective traditions of democracy" (Abril, 2015; Papacharissi, 2004: 267). Under this logic, *incivility* can be understood as the collection of behaviors that can threaten democracy or stereotype any social group. Incivility is consequential; it reduces trust and legitimacy, which are necessary for democratic well-functioning (Papacharissi, 2004; Hmielowski, Hutchens and Cicchirillo, 2014).

As a type of incivility, trolling on Twitter has yet to be explored, with some exceptions (see for example Sonnenbichler and Bazant, 2012). However, unlike this study, Sonnenbichler and Bazant's research was not focused on trolling. Likewise, they fetched tweets from the API and not the firehose. Lastly, there was no information given about the contentiousness of the issues discussed in the hashtags the authors followed. Therefore, the definition of trolls used here is adapted from Herring and colleagues (Herring *et al*., 2002), who analyzed trolls within a controversial context. The Herring study looked at trolling on a message board, which is different from a Twitter discussion in the number of members, reach, and openness of the discussion. However, it is one of the few studies to treat trolls in a meaningful depth. Based on their work, this study conceptualizes Twitter trolls as (a) someone who appears sincere but is actually not sincere; (b) someone posting tweets designed to attract heated discussion and flames; and (c) someone posting tweets wasting participants' time by provoking futile argument and annoying participants (Herring *et al*., 2002: 375).[1]

In contrast to Herring and colleagues (2002), who were able to study a group with well-defined boundaries in terms of topics, the architecture of Twitter is very different. Twitter does not allow for the many strategies forums and chats employ to combat trolls, like excluding users before entrance, banning users after trolling, filtering messages before posting, or disallowing communication between users (Herring *et al.*, 2002). In fact, Twitter's Terms of Use does not contain the word "troll", though they do prohibit abuse of the system and have boundaries regarding what can be posted.

The limited literature on trolling has portrayed trolls as vocational provocateurs, who are apt at steering conversation their way in order to stir conflict. Trolls typically generate messages with racist, xenophobic, homophobic, misogynist, classist, or similar content (Tabachnik, 2012). However, trolls start by enticing audiences in order to get (false) "allies" in the discussion. Answering to the troll's call is called "feeding the trolls" (Tabachnik, 2012). A troll is successful, if users are deceived into believing the troll's intention(s) and are provoked into responding sincerely (Hardaker, 2010). Novices are easy prey for trolls in this sense. Contrary, "a troll with no response has failed" (p. 233). The more responses trolls get, the more successful and amused they are said to be.

One common thread in trolling activity research is the use of foul language (Herring *et al.*, 2002; Tabachnik, 2012; Galán-García *et al.*, 2014; Younus *et al.*, 2014). Though insults do not start at the onset of trolling (otherwise trolling would fail or it would be labeled "flaming"; Hmielowski, Hutchens and Cicchirillo, 2014), when audiences begin biting the bait and futile discussion ensues, insults emerge (Herring *et al.*, 2002). To be sure, flaming does not always mean trolling, and not all trolling contains flaming. The difference between trolling and flaming lies in the motives (Hmielowski, Hutchens and Cicchirillo, 2014): Flaming does not carry deception (Hardaker, 2010), which is an inherent element of trolling. Of equal importance, harassment and bullying are different from trolling (see <www.huffingtonpost.com/news/twitter-trolls>). Harassment against women, minority groups, or celebrities may emerge during trolling.

The interest here lies in exploring the presence of *successful* Twitter trolling activity and in uncovering their main characteristics. Because Twitter discussions are vast, the choice was to select a particular discussion with time and geographical boundaries. One such case was the Catalan independence discussion on Twitter, which took place before the 2012 Catalan Parliamentary election (Guerrero-Solé, Corominas-Murtra and López-González, 2014).

TROLLS ON TWITTER

Social networking sites like Twitter represent a fertile ground for online democracy since —unlike blogs or forums, which never enjoyed more than a fraction of audience-produced content— over 70% of people in the US use social networking sites (Brenner and Smith, 2013), i.e. most of them read or post messages, photos, videos and links. Twitter is the main *public* online platform (Esquire, 2015).[2] There are

few entry barriers, is available in 40+ languages, and its users send over 500 million tweets per day (Twitter, 2016). Moreover, Twitter can act as a catalyst for action or as a vibrant expression of the public sphere (Habermas, 1989) —at least of unparalleled democratic discussion (Huckfeldt, Johnson and Sprague, 2004; Mansbridge, 1999; Mutz, 2006)—.

Twitter's democratic potential stems from its design and architecture. Getting an account is free and easy if consumers are online. Although some few accounts are set up to be private, most accounts are public (the default at sign-up). Therefore, when conversations emerge, they are experienced by a broader audience than the one directly participating in them (Boyd, Golder and Lotan, 2010). Even those without a Twitter account can observe a Twitter conversation; hence, the potential is vast. Numerous scholars have attested to the affordances of Twitter as a discussion space (see Boyd, Golder and Lotan, 2010; Honeycutt and Herring, 2009).

With these affordances, though, there comes a risk; a risk that someone may interrupt the discussion, ridicule it, or even threaten it. The consequences of this can be discouraging since it may shut off participants in the future or prevent new ones from joining in. For instance, trolls can deter civil discussion by spreading distrust, which would have disastrous consequences for the online sphere (Douai and Nofal, 2012) and society at large.

To better understand trolls, this study also seeks to characterize Twitter trolls. Research on personality traits of Twitter account holders who engage in trolling behavior —using metadata or data about user accounts— shows that "psychopathy" and "Machiavellianism" are typical (Sumner *et al.*, 2012). Psychopathy is typified by a lack of empathy and guilt, persuasive speech, pathological lying, a grandiose sense of self-worth, anti-social and promiscuous behavior, and a parasitic lifestyle (Sumner *et al.*, 2012, p. 387). Machiavellianism consists of deceiving and manipulative tendencies toward others, usually for personal gain (p. 387). Similarly, trolls seek to disrupt spaces and attract responses (Herring *et al.*, 2002), which on Twitter may translate into hashtag communities or tweets directed at accounts (people or organizations). Lastly, trolls also hack names to resemble known accounts (Metaxas and Mustafaraj, 2013) or use fake accounts, and tend to have recent accounts (Sumner *et al.*, 2012).

THE CASE OF THE 2012 PARLIAMENTARY ELECTION IN CATALONIA

Catalonia is an autonomous community of Spain with a distinct language, history, government, law, culture, and traditions (Generalitat de Catalunya, no date). Although Catalans have always felt different from Spaniards (Llobera, 1983), recent events have exacerbated this tendency. First, a mounting dissatisfaction among the general population since the 2008 recession. Second, an increased denial of autonomy from the central government in areas such as education and Catalan language use, which culminated in the Supreme Court overturning the 2006 Statue of Autonomy proposal in 2010. Alongside, these developments, younger generations

are being brought into the job market with prospects of over 50% unemployment. Younger Catalans are also a generation now placing more emphasis on Catalan traditions (Serra and Puig, 2012) than their older counterparts. This situational context has created a suitable setting for the independence sentiment to flourish. In 2012, independence supporters outnumbered detractors for the first time in modern history (*El País*, 2014), a statistic that has since continued to flock around 50%.

In 2012, on Catalonia's National Day (September 11), about 1.5 million marched in the streets of Barcelona to defend Catalonia's independence, becoming the most-attended national day in history. The then Catalan government seized that opportunity to end the legislature and run an early election that same year with the intent to win even more seats in congress and lead Catalonia toward independence.[3] The entire focus of that election was Catalonia's independence, forcing other parties with ambiguous independence positions to take a side. The election was set to take place on November 25, 2012, and the campaign to start two weeks prior. However, pre-campaign activities started right after the call and intensified toward the start of the official campaign period —the rationale for four weeks of data—.

In Catalonia, only 9.6% of adults have a Twitter account, yet 15.5% of them obtain information about political debates from Twitter and 29% said they obtained a lot of information from the network (Centre d'Estudis d'Opinió, 2016). In recent elections, Twitter has been used for information seeking, news search, opinion expression, discussion, and polling, precisely because of the importance of Twitter's reach in Spain (Barberá and Rivero, 2015) and in Catalonia (Salcedo, 2013).

Even though Catalonia is a region with about nine million people, Catalan is an official language on Twitter and one of the top 10 languages in terms of internet penetration (Generalitat de Catalunya, 2015). Around elections, discussions on Twitter have been abundant and heated, making Twitter a democratic space for political discussion (Barberá and Rivero, 2015). During the 2012 election campaign, Twitter echoed the heated discussion among Catalans —and between Catalans and Spaniards— about the legitimacy, likelihood, and projection for Catalan independence. Because the independence issue was polarized (Salcedo, 2013), the Catalan independence discussion on Twitter exemplifies a fertile ground for trolling (Conover *et al.*, 2011). Given an election whose main purpose was to gauge the independent movement, in which Twitter echoed the independence discussion and in which the potential for trolls was heightened, the following research question is proposed:

- ***RQ1: What was the extent of successful trolling activity in the Catalan independence discussion on Twitter?***

To better understand trolling behavior, this study also seeks to identify some of the trolls' features. Provided the general presence of Psychopaths and Machiavellianists, hashtag and top account hackers, usernames that resemble known ones or are fake, accounts that are young, and the possibility that more features may emerge in trolling behavior, the following research question is posed:

- ***RQ2: What were the characteristics of successful trolls in the Catalan independence discussion on Twitter?***

METHODS

Data

The population consisted of all tweets generated during the four weeks before the Parliamentary Election in Catalonia using the fetch words "independència" (Catalan for independence), "independencia" (Spanish for independence)[4] or "#25N" (the hashtag for the day of the election), which generated 434,507 units. Although the official campaign spanned two weeks, the discussion had been heated in weeks prior (Gordillo, 2012), so collecting four weeks before Election Day provides sufficient tweets for analysis. By using the GNIP PowerTrack (Twitter Firehose) to capture the tweets, it allowed to include *all* the units in the search. Most Twitter fetches in published research use the API feed —sampling from the search words at a rate not always known, and thus compromising the ability to calculate recall—. Even though some researchers claim that for large datasets like this the API performs nearly as good as the Firehose (Morstatter *et al.*, 2013), data from previous analyses suggests this is not always the case and it can be challenging to predict in advance (Bruns, 2012; Kim, Huang and Emery, 2016).

Among the collected tweets, a computerized content analysis for relevancy uncovered that about 24% of them were irrelevant to the Catalan independence discussion, thus leaving a corpus of 325,888 relevant tweets for analysis. Hence, *precision* (the ability to avoid extraneous tweets) was at 76%, which is below the desirable goal of about 90% (Stryker *et al.*, 2006). Nevertheless, it would have been problematic to achieve better precision given that, at the same time of the Catalan Parliamentary election, there was the general election in the United States, and the peoples of Puerto Rico were voting on their own independence. Precision was calculated via human coding of a sample (n = 500), and then letting the system (Texifter) learn and machine code the rest. The intercoder reliability for the human coding was Kappa = .91, which is acceptable (Landis and Koch, 1977).

The ability to accurately retrieve items of interest as discussed in Stryker and colleagues (2006) —what is called *recall*— could not be assessed because the researcher did not test additional words that could imply independence without using the actual "independence" word. Therefore, the exact population of tweets *relevant* to the independence discussion (without using the fetch words) is unknown.

ANALYSIS

Trolling activity is, by definition, deceptive (Donath, 1999; Herring *et al.*, 2002), but deception in an online environment is extremely problematic to capture —even with computational linguistics since there is no single cue for deception (Hirschberg, 2010)—. To infer latent meaning such as deception, researchers require sophisticated computational techniques like latent semantic analysis (Kireyev, Palen and Anderson, 2009) or latent topic models (Xu *et al.*, 2011). Some scholars have been able to detect deception in online dating profiles (Toma and Hancock, 2012), but they described their deception as "high stakes", while the

deception for the independence discussion had lower stakes. Moreover, Toma and Hancock were tasked with searching for deception in a corpus of units for which there was a structure (online dating profiles), whereas the tweets about the independence discussion do not have such structure.

Therefore, this analysis takes an indirect route to finding successful trolling activity: The researcher captured trolls using profanity. This is not to say profanity equals trolling. But *successful* trolling, most of the time, contains traces of profanity in the thread. Two strands of research support this choice. First, psychopathy and Machiavellianism (Sumner *et al.*, 2012) —characteristics of Twitter trolls— lend themselves to profanity (Sumner *et al.*, 2012). Second, research on trolling in other online communities, such as public forums or discussion boards, notes that trolls utilize profanity when discussion fades (Donath, 1999) or after the "victim" baits (Herring *et al.*, 2002; Golder and Donath, 2004). Basically, foul language is typical for successful trolls (Herring *et al.*, 2002; Tabachnik, 2012; Galán-García *et al.*, 2014; Younus *et al.*, 2014).

To capitalize on this characteristic of trolling behavior, tweets with profanity were extracted.[5] This yielded 1,972 tweets, the profanity population (0.61% of the relevant corpus). Minor insults or swear words did not generate a considerable list of tweets, contrary to what happened with stronger swear words.

Since not all profanity denotes trolling behavior, further analysis is warranted. A characteristic of trolls is targeted behavior, not mass-oriented (Golder and Donath, 2004). Trolls have a group of people or individual in mind when they attack. Therefore, I argue that trolling behavior ought to contain a popular hashtag (#) to enter a particular discussion space or an "@" mention to be directed at an account (individual or group)—or both. Filtering the set with these criteria yielded 1,275 tweets or 0.39% of the corpus.

Furthermore, retweets (RT) and modified tweets (MT) were eliminated because they do not correlate with Machiavellianism or psychopathy (Sumner *et al.*, 2012). In total, 491 tweets (about 25% of the tweets with profanity and 0.15% of the corpus) constituted the set of messages for analysis of potential successful trolling activity.

To detect *successful trolling* activity among the 491 tweets, manual coding was conducted based on the following two criteria: First, successful trolling activity had to at least generate a response; a conversation. Otherwise, they remain unbitten bait or flaming. The response from the "victim," was also analyzed to see whether they killed the troll or bit it. Second, and based on the study's definition of Twitter trolling, trolling activity must entail a conversation (Herring *et al.*, 2002) in which the troll exhibits any of the following behaviors: manifestation of sincerity, flame bait, attempts to provoke futile arguments, attempts to annoy, and ideological manipulation. In order to do this, each one of the tweets will be expanded to read the entire conversation stream if present. Overall, this analysis will illustrate the trolling path from laying bait to successful trolling, in which a string of responses will exhibit the "victim" biting the bait.

In order to analyze who the trolls are, the analysis of the potential 491 trolls operationalized the following features. (a) Psychopaths and Machiavellians tend to

have higher *Klout* (Sumner *et al.*, 2012). (b) Trolls try to *hack hashtags* to become trending topics (Recuero, Amaral and Monteiro, 2012). (a) Because trolls are designed to attract responses (Herring *et al.*, 2002), they may most likely be *directed at known accounts* (i.e., accounts from stakeholders in the conversation). (d) Trolls tend to have *user names* that resemble those of key players in a discussion like political parties, media figures (Metaxas and Mustafaraj, 2013), or simply fake account names (Rafferty, 2011). Finally, (e) since most trolls tend to create accounts shortly before they start trolling, the search for trolls will be centered on *younger* accounts,[6] that is, relatively new accounts created shortly before the election (Sumner *et al.*, 2012).

RESULTS

In responding to whether there were any trolls in the Twitter independence discussion (RQ1), the behavior of the 491 potential trolls[7] was analyzed. Of these, only 105 (21.39%) generated any response. Of these 105 tweets with potential trolling behavior, only 23 were actual successful trolls. The intercoder reliability for the human coding was Kappa = 1 in the potential trolling set, which is acceptable (Landis and Koch, 1977). To double-check the likelihood of capturing successful trolls was valid and reliable, a new random sample of the Twitter independence discussion (non-profanity population; $n = 200$) was selected for further analysis (human coding) of trolling behavior. Each of the sample tweets was read, together with the discussion in which they were embedded (if so). Within this sample, no successful trolls were found. The intercoder reliability for this random sample was Kappa = .96, also acceptable (Landis and Koch, 1977).

Below is an example of a successful troll[8] (account names are blinded):

> A
>
> Lo de Toni Cantó es lamentable, además de ser un actor mediocre, se ríe del nacionalismo catalán, Unión, SI Progreso y Democracia? NO [That about Toni Cantó is unfortunate, besides being a mediocre actor, he laughs at Catalan nationalism, Union, YES Progress and Democracy? NO]
>
> B
>
> @A qué poco humor tenéis [You have little humor]
>
> A
>
> @B humor? Eso es reírse de miles de personas que manifiestan pacíficamente sus ideas independentistas. [Humor? This is actually laughing at thousands of people who peacefully protest their independence ideas.]
>
> B
>
> @A si es que es para reírse, eso de la independencia es cosa de risa. Cuándo fue Cataluña independiente? Todo es un mamoneo [It is totally laughable, that independence thing is a joke. when was Catalonia independent? it sucks]
>
> A
>
> @B eres un ignorante, Cataluña desde siempre ha tenido sentimiento independentista, yo con

fachas como tú, no hablo. A mamarla ;) [You're an ignorant, Catalonia has always had an independence sentiment, with fascists like you, I do not speak. Suck it up;)

B

@A ignorante eres tú, subnormal, facha? Pero qué perro de mierda eres si te tengo delante ...más fachas que vosotros nadie perro [the ignorant is you, retarded, fascist? but what kind of a dog are you if I were in front of you… Nobody is more fascist than you bitch]

A

@B si estoy delante tuyo qué?? Así sois los fascistas siempre amenazando con violencia jaja que payaso eres chaval [if I'm in front of you so what? Fascists like you are always threatening with violence ha ha what clown you are, boy]

***B

@A estás trastornado, no me extraña que queráis la independencia, solo veis enemigos, fachas, fascistas... joder el retraso te corroe [you're deranged, no wonder you want independence, you only see enemies, fascists, fascist ... fuck your retardness is gnawing you]

A

@B Eee para el carro, yo no soy independentista, no soy ni catalán, soy navarro y también me siento español, pero… [hey, hold your horses, I am not independentist, I am not Catalan, I am from Navarra and I also feel Spanish, but…]

A

@B no soy un español que pierde el culo por El Rey y grita Viva España como un baboso sin sentido. Esa España no, esa es de los ricos [I am not a Spaniard who is losing his ass for The King and shouts Long Live Spain like a drooler without any sense. This Spain no, this is the one for the rich]

A

@B y de los capitalistas manipuladores que se lucran a nuestra costa [and the one that belongs to manipulative capitalists who profit at our expense]

B

@A ya vale de lamentaciones, se sabe que tenemos chupones en todos los sitios, lo peor d todo es que se aprovechan para remover... [enough about lamenting, we all know we have money suckers everywhere, the worst is that they take advantage to remove..]

Within the potential trolling set, on several occasions, trolling activity ended up with participants in the conversation apologizing to each other for the misunderstanding, thus killing the troll. That would be an example of an unsuccessful troll. For instance:

A

què deia la manifestació de l'11S? Catalunya nou Estat d'Europa ... l'enquesta ho corrobora #25N #dbtcataluña [what did the 11S protest say? Catalonia the new European state…. The survey supports this #25N #dbtcataluña]

***B

@A habla español que no se te entiende, que sus vais a joder, de independencia nasti de

plasti [speak Spanish otherwise you cannot be understood, you are going to be fucked, no nee nee independence]

A

@B ui es que no me dirigía a ti, perdona, ya lo haré en castellano. [oh I wasn't talking to you, sorry, I will do it in Spanish .]

C

@A @B eres la mar de amable verdad? Que educación chico! [you are so polite, right? These are manners, boy!

B

@A mil disculpas, lamento lo sucedido, creo que me equivoque al tweettear, saludos y disculpas [my apologies, I am so sorry for what happened, I think I made a mistake tweeting, greetings and apologies]

A

@B comprensible. Buenas noches [understandable. Good night]

It is also important to add as a follow-up to RQ1 that results showed most statements of disagreement were not part of any trolling activity. There may be more potential trolls than this study has found —since the analysis was designed to capture successful ones— but I suspect there would not be much more successful trolling activity than what has been reported here. From the successful trolling set (n = 23), most trolling activities used indirect attacks, innuendo, or insinuations rather than direct attacks or accusations.

Table 1. Comparisons across Groups

Variable	Corpus	Trolling
Klout score	49.13	38.19
Top hashtags over all hashtags	49.435%	45.24%
Top user mentions over all mentions	13.86%	41.37%
Fake accounts		64.29%
N	325,888	23

To know who the trolls were and what their trolling pattern was (RQ2), an analysis of the characteristics of the corpus of tweets and the 23 instances of successful trolling activity was conducted (collected in Table 1). In comparison, accounts with trolling activity did *not* have a higher Klout than the general overall corpus, contrary to Sumner *et al.*'s results (2012). They did *not* use more top hashtags over all hashtags (Recuero, Amaral and Monteiro, 2012) either. However, these trolling accounts used more top mentions over all mentions (Herring *et al.*, 2002) than the entire corpus.

About 64.36% of the trolling tweets came from fake name accounts or concealed accounts. The account age of the trolling tweets was checked and none of them were younger than a month counting from Election Day. The oldest account was a bit over five years, while the youngest was two months.

The analysis also revealed that there were accounts created exclusively to troll; but trolling was also an occurrence in genuine accounts. Numerous accounts with an explicit agenda (e.g., anti-Catalan, anti-Spain, pro-Madrid, anti-corruption) were also found among the successful trolling set. Yet, since their motives were always displayed, there was little trolling activity taking place on their part. Troll attacks to these accounts went unbitten.

In conclusion, most of the independence discussion was captured (precision was 76%), containing 325,888 relevant tweets. Out of these, 1,972 profanity tweets were spotted and content analyzed. Yet, only 23 tweets generated a successful trolling response. This means that out of the corpus of relevant tweets analyzed, only 0.01% constituted trolling activity, a substantially low amount.

DISCUSSION

This study sought to assess the existence and behavior pattern of successful trolls for the Twitter discussion of the Catalan independence during the 2012 Parliamentary election. The first research question dealt with the existence of successful trolling activity. Results showed that there was some successful trolling activity as exemplified in the 23 tweets, but these only represented 0.01% of the corpus. Fortunately, this result does not represent a remarkable amount and it mimics the findings by Sonnenbichler and Bazant (2012), who did not find trolls in their analysis of hashtag communities and those of Papacharissi (2004) who found that incivility and impoliteness did not dominate online discourse.

A lingering question is why trolls have generated such media fuss (Delclós Juanola, 2013; Mooney, 2014). One explanation could be that even one successful troll in a meaningful discussion can have negative consequences (e.g., see Herring *et al.*, 2002). These consequences could be then amplified and diffused in the mass media. Similarly, the effect of trolls and other incivility such as flaming can be damaging, even in low amounts (Hmielowski, Hutchens and Cicchirillo, 2014). Still, understanding the underlying infamy of trolls in the media is an aspect that only future research can tackle.

Thinking about how to capture trolls, the present study noted that analysis of a non-profanity random set yielded no additional trolls, thus providing support for the methodology employed: Catching potential trolls via profanity and a few rules that can be computerized (like the presence of mentions, hashtags, or the text (tweet) being embedded in a conversation) is effective. However, this method does not provide a set of baiting or potential tweets that are not successful (i.e., not responded to but that still could be damaging to democratic discussion). Let's remember that, at the baiting stage, profanity is not yet employed, and so these tweets were not captured by this study. But detecting potential trolling tweets is difficult because deception (used in the baiting stage) cannot be detected with machine coding (Hirschberg, 2010).

The second research question inquired about the characteristics of these trolls. The analysis of the 23 successful trolling tweets revealed that trolling account holders (trolls) did not have a higher than average Klout (Sumner *et al.*, 2012), or ratio

of top hashtags over all hashtags (Recuero, Amaral and Monteiro, 2012). Their ratio of top user mentions over all mentions (Herring *et al.*, 2002), on the other hand, was higher than in the general corpus of tweets. This result together with the low incidence of successful trolling activity may indicate either an apprehension to conflictive discussion on Twitter—the potential trolls accomplished silencing a democratic discussion—or an aversion to trolling behavior. The former is not desirable for democracy, but the latter may be advantageous for democracy if uncivil discussion is ultimately shut down. After all, disruptive messages can have severe consequences for the deliberative system in democratic societies (Herring *et al.*, 2002; Hmielowski, Hutchens and Cicchirillo, 2014) and could frustrate the development of a public sphere. However, this is a question that only a study asking Twitter users can answer, and which lies outside of the this study's purview.

Some limitations in this study should be noted. First, recall could not be assessed because fetching tweets that discussed the independence aspect of the 2012 Parliamentary election without using the word independence or the hashtag #25N was difficult. However, the corpus of 325,888 should provide significant insight into the main conversational and trolling-related aspects of the discussion. Second, there may be successful trolling behavior without profanity. However, most literature consulted indicated profanity shows up in trolling behavior, either on the troll or the "victim" side (Donath, 1999; Herring *et al.*, 2002; Golder and Donath, 2004; Sumner *et al.*, 2012; Hmielowski, Hutchens and Cicchirillo, 2014). Moreover, the additional analysis outside of the profanity set showed that there were not additional trolls.

Despite these limitations, this study makes a positive contribution to the analysis of incivility on Twitter and trolling, in particular, for a contentious local issue: the Catalan independence discussion. Study analyses take an empirical look at a naturally-occurring setting (i.e., not in the lab), which provides ecological validity by working with a population of tweets rather than a sample. Further, this study advances a new procedure for identifying trolls through profanity, which may serve future studies on Twitter or other online platforms identify this type of incivility. Other types of incivility, however, remain unexplored here.

Lastly, the results here underscore a Twittersphere in which users decide to solve their discussion viewpoints without resorting to trolls. Thus, it so seems like the discussion was closer to the democratic ideal of an online conversational space (Honeycutt and Herring, 2009).

Eulàlia P. Abril has a Ph. D. from the University of Wisconsin-Madison and is currently an assistant professor of Communication at the University of Illinois at Chicago. Her research areas include health and political communication, looking at new media effects, persuasion, and public opinion. She has chaired the Midwest Association for Public Opinion Research Annual Conference and co-chaired the Mobile Preconference at the International Communication Association.

Notes

1 Hardaker (2010: 237) has a similar definition. Yet she did not quantify types of trolling activity (successful versus not).

2 Only behind Facebook, but Facebook's capacity to connect to the offline world (like TV news) or Facebook's openness is second to Twitter.

3 Though this assertion remains a contentious subject.

4 Twitter developers (dev.twitter.com) pose that fetches without accents usually bring both the word with and without accents; however, to be sure, I included both terms to warrant a complete fetch. I could not find a source that would *guarantee* that searching without accents would fetch both words.

5 The most used insults and their variants from <http://www.taringa.net/posts/humor/5029775/Listado-de-insultos-en-espanol.html> and <http://webs.racocatala.cat/cat1714/insults.htm> were used.

6 Using <http://twbirthday.com/>.

7 Three of these 491 tweets were removed afterwards by their account holders, or the account was deleted.

8 *** Indicates the tweet captured with the search.

References

Abril, E. P. (2015). "Can Partisan News Be Valuable for Discussion? An Analysis of the Effects of Internal Balance on Online Discussion Intention". *International Journal of Communication,* 9, pp. 1029-1051.

Abril, E., P. and Rojas, H. (2015). "Silencing Political Opinions in a Post-Electoral Context." *Communication Research*. Advance online publication. <doi: 10.1177/0093650215616455>

Abril, E. P.; Szczypka, G.; and Emery, S. L. (2017). "LMFAO! Humor As a Reaction to Fear. Decomposing Fear Control within the Extended Parallel Process Model". *Journal of Broadcasting and Electronic Media*. Advance online publication. <doi: 10.1080/08838151.2016.1273921>

Barberá, P. and Rivero, G. (2015). "Understanding the Political Representativeness of Twitter Users". *Social Science Computer Review*, 33(6), pp. 712-729. doi: 10.1177/0894439314558836.

Boyd, D.; Golder, S. A. and Lotan, G. (2010). "Tweet Tweet Retweet: Conversational Aspects of Retweeting on Twitter". In: *IEEE Second International Conference on Social Computing*.

Brenner, J. and Smith, A. (2013). *72% of Online Adults Are Social Networking Site Users*. Techreport. Washington, D.C. Available at: <http://pewinternet.org>.

Bruns, A. (2012). "How Long Is a Tweet? Mapping Dynamic Conversation Networks on Twitter Using Gawk and Gephi, Information". *Information, Communication & Society,* 15(9), pp. 1323-1351. doi: 10.1080/1369118X.2011.635214.

Centre d'Estudis d'Opinió (2016). *Xarxes socials i usos educatius*. Barcelona, Spain. Available at: <http://mestres.ara.cat/delasocietatdigitalalesaules/2011/05/18/xarxes-socials-i-usos-educatius/comment-page-1/>.

Conover, M. D.; Ratkiewicz, J.; Francisco, M.; Gonc, B.; Flammini, A. and Menczer, F. (2011). "Political Polarization on Twitter". In: *Proceedings of the 25th Conference of the Association for the Advancement of Artificial Intelligence*.

Delclós Juanola, T. (2013). "El desalojo de 'trolls'". *El País,* 24 November. Available

at: <http://elpais.com/elpais/2013/11/22/opinion/1385151021_264923.html>.

Donath, J. S. (1999). "Identity and Deception in the Virtual Community". In: Smith, M. A. and Kollock, P. (eds.) *Communities in Cyberspace*. London: Routledge, pp. 25-59.

Douai, A. and Nofal, H. K. (2012). "Commenting in the Online Arab Public Sphere: Debating the Swiss Minaret Ban and the 'Ground Zero Mosque' Online". *Journal of Computer-Mediated Communication,* 17(3), pp. 266-282. <doi/10.1111/j.1083-6101.2012.01573.x/full>.

Doval Avendaño, M. and Martínez Rodríguez, B. (2012). "La audiencia activa en Twitter: Análisis de la retirada de un artículo de opinión en *El Mundo*". *Estudios Sobre El Mensaje Periodistico*, 18(1), pp. 55–71.

El País (2014). "Encuesta sobre el contexto político en Cataluña". *El País,* 19 December. Available at: <http://elpais.com/elpais/2014/12/19/media/1419007798_218610.html>.

Emery, S. L.; Szczypka, G.; Abril, E. P.; Kim, Y. and James, L. (2014). "Are You Scared Yet? Evaluating Fear Appeal Messages in Tweets about the Tips Campaign". *Journal of Communication,* 64(2), pp. 278-295.

Esquire, S. B. (2015). *The Moz Top 500, The Moz Top 500*. Available at: <http://moz.com/top500>.

Galán-García, P.; Gaviria de la Puerta, J.; Laorden Gómez, C.; Santos, I. and García, P. B. (2014). "Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying". In: *International Joint Conference SOCO'13-CISIS'13-ICEUTE'13. Advances in Intelligent Systems and Computing*. Salamanca, Spain, pp. 419428. doi: 10.1007/978-3-319-01854-6_43.

Geiger, R. S. (2016). "Bot-Based Collective Blocklists in Twitter: The Counterpublic Moderation of Harassment in a Networked Public Space". *Information Communication & Society*, 19(6, SI), pp. 787-803. doi: 10.1080/1369118X.2016.1153700.

Generalitat de Catalunya (2015). *10 Reasons for Digital Products to Speak Catalan*. Barcelona, Spain. Available at: <http://llengua.gencat.cat/web/.content/documents/publicacions/altres/arxius/10_raons_ang.pdf>.

Generalitat de Catalunya. *Estatut d'autonomia de Catalunya (Statute of Autonomy of Catalonia)*.

Golder, S. A. and Donath, J. (2004). "Social Roles in Electronic Communities". In: *Association of Internet Researchers Conference Internet Research 5.0,* pp. 1-25.

Gordillo, S. (2012). "Twitter ja està en campanya electoral a Catalunya". *El Periódico,* 6 November. Available at: <http://www.elperiodico.cat/ca/noticias/eleccions-2012/twitter-campanya-eleccions-catalunya-2243401#>.

Guerrero-Solé, F.; Corominas-Murtra, B. and López-González, H. (2014). "Pacts with Twitter. Predicting Voters' Indecision and Preferences for Coalitions in Multiparty Systems". *Information, Communication & Society,* 17(10), pp. 1-18. doi: 10.1080/1369118X.2014.920040.

Habermas, J. (1989). *The Structural Transformation of the Public Sphere*. Cambridge, MA: MIT Press.

Hardaker, C. (2010). "Trolling in Asynchronous Computer-Mediated Communication: From User Discussions to Academic Definitions". *Journal of Politeness Research,* 6(2), pp. 215-242. doi: 10.1515/JPLR.2010.011.

Herring, S.; Job-Sluder, K.; Scheckler, R. and Barab, S. (2002). "Searching for Safety Online: Managing 'Trolling' in a Feminist Forum". *The Information Society.* Taylor & Francis, 18(5), pp. 371-384.

Hirschberg, J. (2010). "Deceptive Speech: Clues from Spoken Language". In: Chen, F. (ed.) *Speech Technology*. Boston, MA: Springer, pp. 79-88. doi: 10.1007/978-0-387-73819-2.

Hmielowski, J. D.; Hutchens, M. J. and Cicchirillo, V. J. (2014). "Living in an Age of Online Incivility: Examining the Conditional Indirect Effects of Online Discussion on Political Flaming". *Information, Communi-*

cation & Society, 17(10), pp. 1196-1211. doi: 10.1080/1369118X.2014.899609.

Honeycutt, C. and Herring, S. C. (2009). "Beyond Microblogging: Conversation and Collaboration Via Twitter". In: *42nd Hawaii International Conference on System Sciences,* pp. 1-10. doi: 10.1109/HICSS.2009.89.

Hosch-Dayican, B.; Amrit, C.; Aarts, K. and Dassen, A. (2014). "How Do Online Citizens Persuade Fellow Voters ? Using Twitter during the 2012 Dutch Parliamentary Election Campaign". *Social Science Computer Review.* doi: 10.1177/0894439314558200.

Huckfeldt, R.; Johnson, P. E. and Sprague, J. (2004). *Political Disagreement: The Survival of Diverse Opinions within Communication Networks*. Cambridge, UK: Cambridge University Press.

Kim, Y.; Huang, J. and Emery, S. L. (2016). "Garbage In, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection". *Journal of Medical Internet Research,* 18(2), p. e41. doi: 10.2196/jmir.4738.

Kireyev, K.; Palen, L. and Anderson, K. M. (2009). "Applications of Topics Models to Analysis of Disaster-Related Twitter Data". In: *NIPS Workshop on Applications for Topic Models: Text and Beyond*.

Landis, J. R. and Koch, G. G. (1977). "The Measurement of Observer Agreement for Categorical Data". *Biometrics,* 33(1), pp. 159-174.

Larsson, A. O. and Moe, H. (2011). "Studying Political Microblogging: Twitter Users in the 2010 Swedish Election Campaign". *New Media & Society,* 14(5), pp. 729-747. doi: 10.1177/1461444811422894.

Llobera, J. R. (1983). "The Idea of Volksgeist in the Formation of Catalan Nationalist Ideology". *Ethnic and Racial Studies*. Taylor & Francis Group, 6(3), pp. 332-350.

Mansbridge, J. (1999). "Everyday Talk in the Deliberative System". In: Macedo, S. (ed.) *Deliberative Politics: Essays on Democracy and Disagreement.* New York: Oxford University Press, pp. 211-242.

Metaxas, P. T. and Mustafaraj, E. (2013). "The Rise and the Fall of a Citizen Reporter".

Mooney, C. (2014). "Internet Trolls Really Are Horrible People". *Slate*, February. Available at: <http://www.slate.com/articles/health_and_science/climate_desk/2014/02/internet_troll_personality_study_machiavellianism_narcissism_psychopathy.html>.

Morstatter, F.; Pfeffer, J.; Liu, H. and Carley, K. M. (2013). "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose". In: *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. Cambridge, MA.

Mutz, D. C. (2006). *Hearing the other Side: Deliberative versus Participatory Democracy.* Cambridge, MA: Cambridge University Press.

Papacharissi, Z. (2004). "Democracy Online: Civility, Politeness, and the Democratic Potential of Online Political Discussion Groups". *New Media & Society,* 6(2), pp. 259-283.

Phillips, W. (2015). *This is Why We Can't Have Nice Things*. Cambridge, MA: MIT Press.

Pont, C., and Capdevila, A. (eds.). (2012). *Del carrer a les urnes: El dret a decidir, en campanya. Comunicació política i comportament electoral a les eleccions catalanes del 2012*. Barcelona: Documenta Universitaria.

Rafferty, R. S. (2011). *Motivations behind Cyber Bullying and Online Aggression: Cyber Sanctions, Dominance, and Trolling Online, Master's Thesis*. JOUR. Ohio University.

Recuero, R.; Amaral, A. and Monteiro, C. (2012). "Fandoms, Trending Topics and Social Capital in Twitter". *Selected Papers of Internet Research.*

Salcedo Maldonado, J. L. (2013). "The Exposure to Political Difference Via Twitter, Analysing the Last Catalan Election". In: *European Political Science Association.* Barcelona, Spain.

Santana, A. D. (2013). "Virtuous or Vitriolic. The Effect of Anonymity on Civility in Online Newspaper Reader Comment Boards". *Journalism Practice*. Routledge, pp. 1-16. doi: 10.1080/17512786.2013.813194.

Serra i Puig, B. (2012). "L'Assemblea Nacional Catalana (ANC): Moviment i política". *Anuari del Conflicte Social,* 1, pp. 528-551.

Sonnenbichler, A. C. and Bazant, C. (2012). "Application of a Community Membership Life Cycle Model on Tag-Based Communities in Twitter". In: *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*. Springer, pp. 301-309.

Stryker, J. E.; Wray, R. J.; Hornik, R. C. and Yanovitzky, I. (2006). "Validation of Database Search Terms for Content Analysis: the Case of Cancer News Coverage". *Journalism & Mass Communication Quarterly.* SAGE Publications, 83(2), pp. 413-430.

Sumner, C.; Byers, A.; Boochever, R. and Park, G. J. (2012). "Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets". In: *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*. IEEE, pp. 386-393.

Tabachnik, S. (2012). "Conversations in Cyber Space. Scope and Limits of the Contractual Hypothesis". *Rétor,* 2(2), pp. 243-259.

Toma, C. L. and Hancock, J. T. (2012). "What Lies Beneath: The Linguistic Traces of Deception in Online Dating Profiles". *Journal of Communication,* 62(1), pp. 78-97. doi: 10.1111/j.1460-2466.2011.01619.x.

Twitter (2016). *Twitter, About Twitter, Inc.* <https://about.twitter.com/company> Xu, J.; Jun, K.; Zhu, X. and Bellmore, A. (2011). "Learning from Bullying Traces in Social Media". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montreal, Canada, pp. 656-666.

Yan, W.; Abril, E. P.; Kyoung, K., and Jing, G. (2016). "Entrapped in One's Blind Spot: Perceptions of Bias in Others and Preparation for Deliberation." *Communication and the Public*, 1(1), pp. 72-90.

Younus, A.; Qureshi, M. A.; Saeed, M.; Touheed, N; O'Riordan, C. and Pasi, G. (2014). "Election Trolling: Analyzing Sentiment in Tweets during Pakistan Elections 2013". In: *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*. Seoul, Korea, pp. 411-412. doi: 10.1145/2567948.2577352.