Sizing Up the Troll: A Quantitative Characterization of Moderator-Identified Trolling in an Online Forum

Mattia Samory
Univ. Padova, Italy
samoryma@dei.unipd.it

Enoch Peserico
Univ. Padova, Italy
enoch@dei.unipd.it

ABSTRACT

A few troublemakers often spoil online environments for everyone else. An extremely disruptive type of abuser is the troll, whose malicious activities are relatively non-obvious, and thus difficult to detect and contain – particularly by automated systems. A growing corpus of qualitative research focuses on trolling, and differentiates it from other forms of abuse; however, its findings are not directly actionable into automated systems. On the other hand, quantitative research uses definitions of "troll" that mostly fail to capture what moderators and users consider trolling. We address this gap by giving a quantitative analysis of posts, conversations, and users, specifically sanctioned for trolling in an online forum. Although trolls (unlike most other abusers) hardly stand out in a conversation e.g. in terms of vocabulary, *how* they interact, rather than *what* they contribute, provides cues of their malicious intent.

ACM Classification Keywords

H.4.3. Information Systems Applications: Communications Applications – Bulletin boards

Author Keywords

Social computing; trolls; online forums.

INTRODUCTION

Moderating online content on a large scale is still an open problem. While automated tools succeed in detecting barefaced forms of abuse, more sophisticated offenders elude even human moderators [20]. This is particularly true of trolls – users who create a context conducive to conflict through subtle use of aggression, deception, and/or manipulation [13]. In fact, trolls are recognized as a widespread and very serious threat to online communities [5].

A growing body of research agrees that the standing notions of online aggression fail to capture trolling behaviour [1, 20]. The motivations behind trolling are still unclear, but self-identifying trolls seem to show the typical traits of antisocial personalities, especially high sadism [4]. Perceived trolls operate covertly, gaining the community's trust before setting out to taunt groups and individuals, and denying responsibility even when proven guilty [13]. To further complicate troll identification, trolling strategies are far less obvious than launching unprovoked attacks: for example, they might use off-the-record communication, or imitate inexperienced users [14]. What is more, trolling posts easily camouflage among the responses they elicit, which are often in their own right abusive or "trolling" [15]. Troll moderation thus requires considerable effort, since it needs unambiguous evidence from several abusive events to frame a troll with sufficient confidence. The line of work we just described is fundamental for framing trolling behaviour and understanding its impact; however, it is mostly qualitative, and provides little *actionable* insight.

On the other hand, quantitative work on automated abuse detection (based e.g. on lexical features and text readability [8], affect [6], post timing, rank, and thread length [17], irrelevance and conflict [10], cognitive load [2], user social connectivity and reputation [23]), does not extract an operational definition of "troll" from specific moderation records. Instead, it uses the term "troll" somewhat arbitrarily to target abuse in varying forms (e.g., opinion manipulation [17] and sockpuppet profiles [12]), or degrees (e.g., detecting banned users [8]).

Taking a large online forum as a case study, we bridge the gap between qualitative and quantitative research on trolls extracting an operational definition of trolling behaviour from specific moderation signals (i.e. we call trolling what the moderators call trolling), and characterize this behaviour through quantitative language and interaction metrics. We begin by detailing these metrics and the data curation process. Then we show that, although automatically identifying trolled threads is relatively easy, accurately pinpointing trolls and trolling posts in such threads is challenging. After a comparative analysis with civil users and other abusers (both over their entire activity on the forum, and specifically when they commit infractions), we show how trolls manage to remain covert while disrupting discussion. We conclude discussing implications for moderation and future work.

TROLLS DEFINED BY MODERATORS

This case study analyses `forum.rpg.net`, a large online forum on roleplaying games sporting more than 50.000 members and 14 million posts in half a million threads, from which we collected data spanning the interval 2000–2014. Since 2012, the forum features a section devoted to public display of moderation actions: when a moderator intervenes against infractions of forum rules, a new ticket in this section reports the indicted

CHI 2017, May 06 - 11, 2017, Denver, CO, USA

ACM 978-1-4503-4655-9/17/05 $15.00
DOI: http://dx.doi.org/10.1145/3025453.3026007

prediction task	#	interaction	text	both
troll posts	147	87 84 91	61 51 71	86 82 89
troll threads	130	93 90 95	58 48 67	92 89 94
troll posts in thread	130	63 60 68	53 43 63	63 57 70
posts by trolls	1000	59 50 68	52 41 62	57 48 66

Table 1. Number of items in the smaller class, and percent accuracy of the four troll detection tasks using interaction features, text features, and a combination of both. Sensitivity and specificity (i.e. correct identification rate of trolls, and of others) follow accuracy values.

user and post(s), along with the accusation and the disciplinary measure taken. We categorize posts as trolling only when the moderator explicitly phrases the accusation accordingly – i.e. when the ticket's text matches `(^|\s)troll`. Out of 1549 infringing posts we recovered, 147 are trolling posts.

It is important to note that this conservative categorization of trolls may still be inaccurate [15]: moderators may misinterpret the intentions of the alleged troll, hold slightly different definitions of trolling, or fail to detect trolling altogether. However, we believe this is the most objective way to capture what the forum actually perceives as trolling behaviour.

Hereafter we refer to users as *civil*, if they do not appear in moderation tickets; *abusers*, if they appear in moderation tickets, but were never explicitly sanctioned for trolling; or *trolls*, if sanctioned at least once for trolling. Civil, abusive, and trolling posts follow the same naming convention.

TEXT AND INTERACTION METRICS

We measure text quality of user posts through metrics of *readability* (using the Automated Readability Index – ARI [21], a score that approximates the US grade level needed to comprehend a passage of text), of *politeness* (through a classifier developed in [9] for assessing civility of a request), and of *thematic coherence* (computed as the cosine similarity of the bag-of-words representation of the post with those preceding in the thread). Moreover, we analyse post *content* matching it against the dictionaries of Linguistic Inquiry and Word Count (LIWC), a software to organize words into psychologically meaningful categories such as "inhibition" or "home" [22]. LIWC is a gold standard in psycholinguistic categorization: although its categories are quite broad and can support many different interpretations, the simultaneous over- or -underrepresentation of *sets* of categories can often provide specific and fairly objective insights. Additionally, we analyse non-verbal behaviour through *interaction* features – e.g. the time of posting, or the number of users in the thread – using the "content-agnostic" feature set proposed in [19] for unmasking post authors.

TROLLED THREADS STAND OUT, TROLLS DO NOT

The goal of this work is to provide insight into the distinguishing features of trolls, rather than to build an accurate moderation system. But as a preliminary step we investigate how difficult it is to detect trolls and trolling posts, and why.

We make this question concrete and quantitative by translating it into four related classification tasks. The first is the "classic" task of distinguishing trolling posts from non-trolling posts. The second is distinguishing posts *in trolled threads* from posts in non-trolled threads – note that even in trolled threads almost all posts (typically all but one) are non-trolling posts. The third task is distinguishing trolling posts from non-trolling posts in the same (trolled) thread. The fourth is distinguishing random posts made by trolls, from posts made by non-trolls – again, note that trolling posts constitute only a small minority of those posts made by trolls.

We use combinations of textual features (LIWC counts and ARI rating), and of interaction features. We perform 10 repetitions of binary, balanced classification (i.e. for each item in the smaller class, we attempt classification with 50% probability of that item, and with 50% probability that of an item chosen randomly without reinsertion from the larger class), using a Random Forest model [3], in a cross-validation scheme. The size of the smaller class and accuracy are reported in Table 1.

The classifier can identify trolling posts and trolled threads with good accuracy (respectively 87% and 93% accuracy). However, identifying trolling posts *within* a trolled thread, and "average" posts by trolls, appears considerably harder (respectively 63% and 59% accuracy). These results suggest that even when seeming to accurately detect trolling posts, the classifier is actually detecting *trolled environments*, rather than trolling posts per se – note that most non-trolling posts are in non-trolled threads. The "average" behaviour of troll users does not appear significantly different from that of other users, and when trolls do act maliciously, they seem to successfully hide within a discussion that ends up uniformly "trollish".

Note that, in all tasks, textual features provide significantly less information than non-textual features, and combining both provides little or no advantage over using interaction features in isolation. In other words, non-verbal behaviour may well be what can actually unmask trolls.

TROLLS ARE RARELY TROLLS

We saw that trolls appear hard to distinguish from other users, both in and out of trolled threads. They do have a few characteristic markers, however. This section focuses on those quantitative differences that set trolls apart from civil users, and from other abusers, *over their entire posting history*.

Trolls are eager, urbane, cold-hearted contributors

It may be surprising that on average trolls contribute to the forum over a timespan of more than 5 years, writing more than 3500 posts – significantly more than civil users, and in line with other abusers (Tukey's test, $p < .05$). Therefore, to avoid artifacts [18], we match each troll to exactly one civil user and one abuser with a similar post rate and total number of posts. This reduces our dataset to a total of roughly 1.2 million posts, authored by 120 users in each category. We then perform a series of 3-way comparisons of text and interaction features between the trolls, abusers, and civil users. All results, unless otherwise stated, are significant by Tukey's test, $p < .05$.

First, we focus on text quality. Trolls write less readable posts, with smaller word count and character count, compared to both abusers and civil users. This may be due to their sacrificing quality for quantity. However, posts by trolls are slightly, but not significantly, more coherent with the 3 preceding posts in the thread than those by civil users ($t = 0.607$, $p = 0.544$). Previous literature found that antisocial users tend to be less coherent than civil users [8]; our findings suggest that trolls attempt to contribute useful content for a large portion of their life to gain the trust of the community [11].

Next, we examine the linguistic choices of users, measuring the frequency of LIWC categories in their posts. Abusers sport stronger use of openly offensive language than trolls and civil users, correlating negatively with "inhibition", "relative", and "social" word categories, and positively with "sexual", "death", "swear", and "bio" (body parts and biological processes) categories. Trolls, instead, just exhibit less empathy and are more confrontational, choosing fewer "inclusive", "positive affect", "future", and "tentative" words, and more "negation" and "causation" words than abusers and civil users.

Even though trolls generally talk more about personal topics (such as "money" or "work") than either civil users or abusers, they talk more than abusers but less than civil users about personal topics with stronger empathic connotations (such as "home"). This is mirrored in the different use of human-related categories. Trolls use less first and third-person pronouns than either civil users or abusers; they use more second-person and first-person-plural pronouns than abusers, but less than civil users. All this suggests trolls are eager to evoke group cohesion [22] (possibly in search for a place in the community, be it honestly or deceptively) but are less able than civil users to sustain it through empathy.

Finally, we look at the different interaction patterns of users. Trolls engage in discussion more eagerly than abusers and civil users (in terms of temporal lag from the start of conversation, number of posts preceding their first post, and propensity to write opening posts). Abusers, on the contrary, are the group with the least propensity to start conversations. Overall, trolls do not quote or get quoted differently from abusers and civil users, but they choose threads with more "intense" interaction: shorter (in terms of number of posts and time between first and last post) but more verbose (in terms of characters per post), attracting fewer views but more views per post, with participants entering the conversation earlier (in terms of number of preceding posts and inter-post lag), and with more pairs of users quoting each other.

What emerges is a profile of the troll as a user that is not *obviously* offensive, asocial or secretive, and that is in fact eager to be part of the community (indeed more than civil users) – albeit somewhat lacking in empathy towards others, and thus harsher, colder, and more confrontational.

Trolls write ever more desperately

We now focus on the changes in quality and quantity of content during the lifetime of trolls, compared to civil users and abusers. To avoid artifacts we match users as in the previous section. We then divide the lifespan (from first to last post) of each user into ten "ages" of equal duration, and compare readability in terms of ARI across user types and ages. Trolls and abusers enter the forum writing less readable text than civil users ($t \approx 8$, $p < 0.001$). All three user types are less readable in the last age than in the first; trolls worsen more than civil users (difference in differences via linear regression, $\beta = .191$, $p < .01$), like abusers.

Civil users see the readability of their posts improve throughout the first half of their lifetime, and slowly worsen in the second. Abusers see it worsen abruptly near the very end of their lifetime. Trolls, instead, produce posts of steadily worsening quality disseminated across an ever increasing number of threads at an ever faster pace (significantly more than civil users or abusers).

The fact that all users see the quality of their posts worsen in their last age may reflect the disaffection that eventually makes them leave the site. The sharp drop in abuser post readability may indicate a well-defined break point, that leads to a sudden departure from social norms; in fact, the majority of infractions happens around this time in an abuser's life. Existing literature confirms that readability of antisocial users starts out lower than that of other users; and it suggests that its subsequent degradation may be may be in retaliation for negative community feedback [7]. While this seems reasonable in the case of abusers, it does not fully explain why trolls would be led to post *more*, and in more threads. In fact, it seems that the steady degradation of troll post readability is the consequence of an unexplained urge to increase their posting rate, sacrificing quality for quantity. In any case, the lack of a sharp change in posting behaviour makes trolls harder to detect than abusers.

PUNISHED MORE HARSHLY, FOR NO OBVIOUS REASON

After examining the "normal" life of trolls, we now focus on their actual trolling behaviour.

We begin by looking at how posts that have been moderated for trolling differ from other moderated posts, using all quality, textual, and interaction features. The language in trolling posts is more controversial (more words in the "sex", "humans" LIWC categories, $p < .05$) than that of other abusive posts. It is, however, not significantly more offensive (e.g. "swear", "negative affect" categories) or incoherent with the previous posts. Trolling posts appear earlier in the thread (in terms of wall clock time), and the conversation preceding the trolling post is more hectic (shorter timespan between posts, more users, and more posts). Overall, trolled threads receive as many replies and views as other abused threads, but in a shorter time, engaging more users, and with more user pairs exchanging quotes. The distinguishing feature of trolling posts, therefore, seems to be the level of excitement that surrounds them, rather than specific language features.

In general, trolling posts are more heavily sanctioned than other forms of abuse, considering the numeric score associated with the gravity of the penalty in the moderation tickets ($t = 2.16$, $p < .01$). However, the "criminal" history of trolls is marked by more infractions overall ($t = 4.29$ $p < .001$), and higher cumulated penalties ($t = 4.32$ $p < .001$). While few users get sanctioned for trolling as their very first post (probably intentionally created sockpuppet accounts), trolls that relapse do not troll as their first infraction. Moderators may require several rounds of sanctioning before correctly recognizing a troll [14], and despite heavier sanctions trolls remain on the site as long as other abusers after the first violation.

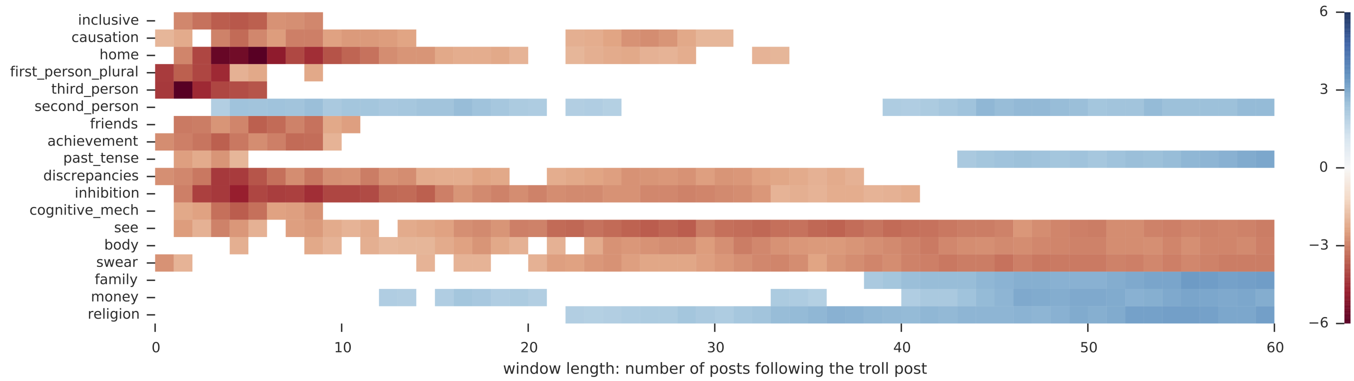

Figure 1. T-test statistic for selected LIWC categories, comparing posts preceding the trolling post to posts following it within a window of growing length. Colours reflects the value of the statistic, ranging from dark red (negative) to dark blue (positive). Only significant results ($p < .05$) are reported.

TROLLED THREADS

This section looks at trolled *threads*, giving context to troll infractions, and the reactions they provoke. All results reported are statistically significant ($p < .05$).

Troll posts: angst and reappraisal

We start by studying the language used by trolling posts, when compared to posts by other users in the same thread. Trolling posts are not obviously insulting (e.g. do not use more words in the "swear" LIWC category), but seem written to induce emotional responses (more "bio", "sex", "anger", "causation", "negative" and "positive affect", "second person" words). Contrary to expectations [11], trolls do not show markers of deception – especially, trolls use complex language (more "exclusions", "prepositions", "cognitive mechanisms", as well as longer text and equal readability) that is supposedly incompatible with the cognitive load that lying requires. Finally, increased use of "causation" and "insight" is associated with reappraisal [22] - trolls may fake reconciliation (as in the case of pseudo-naive trolls [14] and “concern trolls”), or change stance in the argument [13]. In conclusion, trolling posts seem to speak to the emotionality of readers, and while they do not show signs of deception, they may mask subtle dialectic strategies. The LIWC categories associated with trolling posts sketch the troll as a hurt individual, as they find correlation in the literature with reworking of trauma, depression, and unsatisfactory relationships [22].

Reactions to trolls: the damage is already done

Finally, we analyse how the trolled threads evolve around the trolling posts. Posts *following* a trolling post differ from ones *preceding* it in that they feature more words in confrontational categories ("causation", "insight", "negation", "exclusive", "certainty"), and markers of debate ("past tense", and "first person singular", "second person", and "indefinite" pronouns). However, emotional charge and amount of obscene words do not differ significantly. That is to say, trolling posts (that get moderated) do not start the fire, but fan the flame.

We investigate further how the effects of trolling posts propagate across thread. We grow a window of posts following the trolling post, and observe changes in LIWC and interaction features, compared to posts preceding the trolling post. Figure 1 depicts the trends in (.05 significant) t statistics for several LIWC features of interest. For posts closely following the troll post, emotional language, swear words, and sex-related words see use comparable to that in posts preceding the trolling post. However, there is a striking lessening in inhibition and inclusive language ("inclusive", "first person plural", "friends", "home"). Coincidentally, posts also become shorter and come at a slower pace, and a higher fraction of their content is quoted text. This may be an indicator of the cyclical, pointless derailments of discussion generated by trolls [16]. Widening the window of observation further from the trolling post, one can see that users return to swearing less, talking more of sensitive subjects (e.g. "money", "family", "religion"), and less of physiological processes ("body", "see"). Use of second-person pronouns increases both soon after the trolling post (possibly for accusations) and later in the thread (possibly for reappraisal). Trolling posts hide among neighbouring posts, and build upon an existing state of excitement in the discussion, to amplify controversy. Note, however, that it is possible that the "real" trolling posts, the ones that originate the argument, appear earlier in the conversation yet elude moderation.

DISCUSSION AND CONCLUSIONS

Although detecting troll posts may appear relatively easy with “standard” techniques, what is actually easy is separating posts out of trolled threads (as all trolling posts are) from posts out of non-trolled threads (as most non-trolling posts are). Separating trolling posts from other posts *within trolled threads*, and more in general trolls from other users, is significantly harder. Results from this paper suggest an alternative approach: detecting trolled *threads*, integrating longitudinal data from *user history*, and monitoring *reactions* in trolled threads to identify trolling posts. In particular, interaction features perform well in revealing discussions that will eventually be trolled, and reactions to trolling posts follow noticeable linguistic patterns. This new framing for troll detection may be directly applicable with little effort, since existing systems already have the annotated data and the tools at hand.

Past research has often conflated generic abusers with trolls. Given that trolls are both more disruptive and less obviously uncivil, future research should target them as their own, separately defined category. Our work shows promising results for a quantitative characterization of trolls in *one* forum; it would be crucial to validate our findings across different platforms.

REFERENCES

1. Jonathan Bishop. 2014. Representations of 'trolls' in mass media communication: a review of media-texts and moral panics relating to 'internet trolling'. *International Journal of Web Based Communities* 10, 1 (2014), 7. DOI: http://dx.doi.org/10.1504/IJWBC.2014.058384
2. Jonathan Bishop. 2016. Trolling Is Not Just a Art. It Is an Science. In *Handbook of Research on Digital Crime, Cyberspace Security, and Information Assurance*. Number July. IGI Global, Chapter 28, 436–450. DOI: http://dx.doi.org/10.4018/978-1-4666-6324-4.ch028
3. Leo Breiman. 2001. Random forests. *Machine learning* (2001), 5–32. DOI: http://dx.doi.org/10.1023/A:1010933404324
4. Erin E. Buckels, Paul D. Trapnell, and Delroy L. Paulhus. 2014. Trolls just want to have fun. *Personality and Individual Differences* 67 (9 2014), 97–102. DOI: http://dx.doi.org/10.1016/j.paid.2014.01.016
5. Catherine Buni and Soraya Chemaly. 2016. The Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech. (2016). https://goo.gl/mG2jDA
6. Erik Cambria, Praphul Chandra, Avinash Sharma, and Amir Hussain. 2010. Do Not Feel the Trolls. In *Proceedings of the 3rd International Workshop on Social Data on the Web (SDoW 2010)*. http://ceur-ws.org/Vol-664/paper1.pdf
7. Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How community feedback shapes user behavior. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media - ICWSM '14*. 41–50. http://arxiv.org/abs/1405.1429
8. Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media - ICWSM '15*. http://arxiv.org/abs/1504.00680
9. Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *Proceedings of ACL* (2013). http://politeness.mpi-sws.org/
10. Imen Ouled Dlala, Dorra Attiaoui, Arnaud Martin, and Boutheina Ben Yaghlane. 2014. Trolls Identification within an Uncertain Framework. In *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*. IEEE, 1011–1015. DOI: http://dx.doi.org/10.1109/ICTAI.2014.153
11. Judith Donath. 1999. Idenity and Deception in the Virtual Community. *Communities in Cyberspace* (1999), 27–58. DOI: http://dx.doi.org/10.1519/JSC.0b013e3181e4f7a9
12. Patxi Galán-García, José Gaviria De La Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. 2014. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL* 24, 1 (2014), 42–53. DOI: http://dx.doi.org/10.1093/jigpal/jzv048
13. Claire Hardaker. 2010. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research* 6, 2 (2010), 215–242. DOI: http://dx.doi.org/10.1515/JPLR.2010.011
14. Claire Hardaker. 2013. "Uh. . . . not to be nitpicky,,,,,but. . . the past tense of drag is dragged, not drug.": An overview of trolling strategies. *Journal of Language Aggression and Conflict* 1, 1 (2013), 58–86. DOI: http://dx.doi.org/10.1075/jlac.1.1.04har
15. Claire Hardaker. 2015. 'I refuse to respond to this obvious troll': an overview of responses to (perceived) trolling. *Corpora* 10, 2 (8 2015), 201–229. DOI: http://dx.doi.org/10.3366/cor.2015.0074
16. Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for Safety Online: Managing "Trolling" in a Feminist Forum. *The Information Society* 18, 5 (2002), 371–384. DOI: http://dx.doi.org/10.1080/01972240290108186
17. Todor Mihaylov and Preslav Nakov. 2016. Hunting for Troll Comments in News Community Forums. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2016), 399–405. http://anthology.aclweb.org/P16-2065
18. Paul R. Rosenbaum and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55. DOI: http://dx.doi.org/10.1093/biomet/70.1.41
19. Mattia Samory and Enoch Peserico. 2016. Content attribution ignoring content. In *Proceedings of the 8th ACM Conference on Web Science - WebSci '16*. ACM Press, New York, New York, USA, 233–243. DOI: http://dx.doi.org/10.1145/2908131.2908156
20. Pnina Shachaf and Noriko Hara. 2010. Beyond vandalism: Wikipedia trolls. *Journal of Information Science* 36, 3 (2010), 357–370. DOI: http://dx.doi.org/10.1177/0165551510365390
21. Edgar A. Smith and R J. Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (U.S.)* (5 1967), 1–14. http://www.dtic.mil/dtic/tr/fulltext/u2/667273.pdf
22. Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54. DOI: http://dx.doi.org/10.1177/0261927X09351676
23. Hyeongseok Wi and Wonjae Lee. 2014. The norm of normlessness: Structural Correlates of A Trolling Community. In *Proceedings of the 2014 ACM conference on Web science - WebSci '14*. ACM Press, New York, New York, USA, 275–276. DOI: http://dx.doi.org/10.1145/2615569.2615663