Legal issues related to the use of Twitter data in language research

Kamocki, Pawel; Hannesschläger, Vanessa; Hoorn, Esther; Kelli, Aleksei; Kupietz, Marc ...

Monachini, Monica; Eskevich, Maria

2021-09-27

http://hdl.handle.net/10138/343041

Kamocki, P, Hannesschläger, V, Hoorn, E, Kelli, A, Kupietz, M, Lindén, K & Puksas, A 2021, Legal issues related to the use of Twitter data in language research. in M Monachini & M Eskevich (eds), CLARIN Annual Conference Proceedings 2021. CLARIN Annual Conference Proceedings, CLARIN ERIC, Utrecht, pp. 150-153, CLARIN Annual Conference , 27/09/2021.

Legal Issues Related to the use of Twitter Data in Language Research

Pawel Kamocki
IDS Mannheim, Germany
kamocki@ids-mannheim.de

Vanessa Hannesschläger
OeAW, Austria
vanessa.hannesschlaeger@oeaw.ac.at

Esther Hoorn
Rijksuniversiteit Groningen, the Netherlands
e.hoorn@rug.nl

Aleksei Kelli
University of Tartu, Estonia
aleksei.kelli@ut.ee

Marc Kupietz
IDS Mannheim, Germany
kupietz@ids-mannheim.de

Krister Lindén
University of Helsinki, Finland
krister.linden@helsinki.fi

Andrius Puksas
Mykolas Romeris University, Lithuania
andrius_puksas@mruni.eu

Abstract

Twitter data is used in a wide variety of research disciplines in Social Sciences and Humanities. Although most Twitter data is publicly available, its re-use and sharing raise many legal questions related to intellectual property and personal data protection. Moreover, the use of Twitter and its content is subject to the Terms of Service, which also regulate re-use and sharing. This extended abstract provides a brief analysis of these issues and introduces the new Academic Research product track, which enables authorized researchers to access Twitter API on a preferential basis.

1 Introduction

Social media data is useful for a wide variety of research disciplines in Social Sciences and Humanities, such as sociology, computer science, media and communication, political science, and engineering, to name a few. Twitter is still one of the most popular platforms for academic research on social media data (see Ahmed 2019). Tweet corpora are also used in linguistics (for example, a tweet sub-corpus is being added to the German Reference Corpus, DeReKo), albeit few tweet corpora are widely known, which may be because due to legal grey areas, such corpora are rarely shared.

Indeed, although most Twitter data is publicly available, its re-use and sharing (especially in a way compatible with Open Science requirements) raise many legal questions related to intellectual property and personal data protection.

2 Intellectual Property perspective on Twitter data

Copyright is an intellectual property right (IPR) that grants the author moral and economic rights over his or her work, including the exclusive right to copy it and make it available to the public.

A work is protected by copyright if it is original, i.e., constitutes the author's own intellectual creation. Although it varies from one jurisdiction to another, very short works such as titles or slogans are often considered unoriginal, as the intellectual creation cannot manifest itself in a very short format. The maximum length of a tweet is currently set at 280 characters (increased from 140 in November 2017), i.e. about 50-60 words in English, which seems more than enough to potentially qualify for copyright

protection. However, in practice, very few tweets reach the maximum threshold, and most of them are considerably shorter: the most common length of a tweet in English has been reported to be only 33 characters long (Perez, 2017), i.e. approximately 6-7 words. Nevertheless, it appears that even works of such short length may still qualify for copyright protection: Kamocki (2020) argues that only n-grams that are no longer than 3 words can safely be regarded as copyright-free.

This does not mean that all tweets are indeed original and protected by copyright. Arguably, in reality, and from the quantitative perspective most tweets (like "Big win!", "LewanGOALski!!!!!!!!!!1111" or "This is crazy LOL") fail to meet the originality criterion. However, a pack of several thousand tweets is likely to contain copyright-protected material (even if it does not include photographs or other media). Therefore, while analysing Twitter data (which necessarily involves at least reproduction, i.e. copying of tweets, an act restricted by copyright law), copyright issues cannot be ignored.

Re-use of copyright-protected content is only legally possible if it is authorised by the author (directly or indirectly) or exempted from authorisation by a statutory exception. The recent EU Directive 2019/790 on Copyright in the Digital Single Market (DSM Directive) introduces a specific framework for text and data mining, including an exception for TDM for research purposes (Article 3). This exception allows research organisations to make copies of the content that they have lawful access to. When it comes to publicly available tweets, the criterion of lawful access is met, as per Recital 14 of the DSM Directive *("Lawful access should also cover access to content that is freely available online"*).

The copies made under the "TDM for scientific research" exception have to be stored with the appropriate level of security, but they can be retained for re-use in other projects or for evaluation purposes. However, the exception does not seem to allow any sharing of the data, although there might be slight variations between implementations in the various EU Member States (for example, the German implementation allows for sharing 'with a specifically limited circle of persons for joint research').

Another intellectual property right that could potentially apply to Twitter is the *sui generis* database right. Under this framework, Twitter could claim an exclusive right in its database of tweets, enabling the company to prevent users from extracting and/or re-using the whole database, or a substantial part thereof, independently from any copyright in the content. Although rarely discussed, this right could considerably limit access to tweets or web content in general. However, it is essential to keep in mind that the *sui generis* database right only applies to companies "formed in accordance with the law of a [EU] Member State and having their registered office, central administration or principal place of business within the [European Union]" (Article 8.2 of the Directive 96/9 on Databases). For companies that, like Twitter, only have registered offices in the European Union (Twitter currently has offices in Dublin, Paris, Berlin, Brussels and Madrid), their operations must be "genuinely linked on an ongoing basis with the economy of a Member State". It is not clear whether this is the case with Twitter. And even if it is, a sample of tweets that can be used in a language research project is quite unlikely to constitute a 'substantial part' of all tweets. In light of the above, the impact of the *sui generis* database right on the re-use of tweets for language research is probably minimal and can be ignored.

3 Contract law perspective on Twitter data

To post tweets, one needs to create a Twitter account and accept (among other documents) Twitter's Terms of Service (ToS)[1]. As per Paragraph 3 of the ToS, although the user retains copyright in his or her tweets, he or she grants Twitter a very broad license to re-use them for free on a non-exclusive basis. This means that someone who would like to copy and share tweets can receive the necessary authorization either directly from the user (which in most cases is unworkable in practice, given the sheer number of Twitter users) or from Twitter. Theoretically, nothing prevents users from re-publishing their tweets outside of Twitter, including, e.g. in .xml format and under an open license.

Twitter ToS also grant every user access to the tweets, although certain actions are expressly forbidden. These include accessing or searching (or attempting to access or search) Twitter content by any means other than interfaces provided by Twitter and scraping tweets without prior consent from the company.

[1] Available at: https://twitter.com/en/tos#intlTerms (access: 27.4.2021).

It seems, therefore, that mining of tweets without specific permission (e.g. without being granted access to the mining interface provided by Twitter), even if done for research purposes, would violate Twitter ToS, which may lead to suspension or termination of the user account(s) that is (are) at the origin of these actions. This might be the reason why those researchers who have indeed scraped data from Twitter may not be transparent about it.

Interestingly, as regards copyright, text and data mining for research purposes by research organisations is covered by the abovementioned exception of Article 3 of the DSM Directive. Article 7 of the same Directive expressly states that any contractual provision contrary to this exception should be unenforceable. In specific contexts, national contract laws (e.g. regulating unfair contractual terms) could also have an impact.

It is far from clear, however, how this will work in practice. In our opinion, if a user (affiliated with a research organisation) scrapes or attempts to scrape tweets for scientific research purposes without specific permission from Twitter, he or she would still violate Twitter ToS (and likely see his account closed or at least suspended), even though he or she would not be liable for copyright infringement. He or she would also be able to retain the copies, according to Article 3 of the DSM Directive. However, Twitter could potentially sue the person for damages for breach of contract. Still, in our opinion, this is quite unlikely to happen taking into account the above-mentioned copyright exception and the limited amount of damages that could possibly be obtained, as well as the hypothetical reputational loss for Twitter for suing a researcher or a research institution. Even if a researcher were sued by Twitter, and found guilty of copyright infringement, the penalty (given the nature of academic research) will likely be moderate (probably not exceeding 10000 EUR); however, the consequences in the relations with funding agencies, and within the research institutions, can potentially be more dire.

4 Technological Protection Measures

Although the authors have not tested it, it can be assumed that scraping tweets is not only in principle forbidden by Twitter ToS but also made impossible (or at least very difficult) by technological protection measures (TPM). In addition to being in principle forbidden by law (Article 6 of the Directive 2001/29 on Information Society), circumvention of TPM is also expressly prohibited by Twitter ToS (and will likely lead to a lifetime ban).

Can the above-mentioned exception of Article 3 of the DSM Directive be interpreted as allowing circumvention of TPM in the context of text and data mining for scientific research purposes? This seems to be the most significant grey area of the new exception, as Article 3 allows rightholders to apply TPM only to the extent necessary to ensure the security and integrity of their networks and databases. In our opinion, Twitter would have an excellent chance to succeed in arguing that TPMs implemented to prevent unauthorised scraping are, in fact, necessary to achieve these goals. However, it remains to be seen how this issue will be worked out in practice.

5 Data protection perspective

In addition to being potentially copyright-protected, tweets should also be regarded as personal data (Gold, 2020), as they contain identifying information (at the very least the user ID, but possibly also location metadata or other identifying content). Therefore, their processing needs to follow the GDPR (even though Twitter is an American company -- as per its Article 3.2, the GDPR applies to foreign companies which offer services to EU citizens), as well as applicable ethical rules.

Twitter provides its users with the possibility to fine-tune their privacy settings, including public availability of their tweets and profile information, which potentially may be interpreted as granting/withdrawing consent. Today, it seems that an average Twitter user should be aware that public tweets can be used for research purposes (Twitter expressly informs their users, in its Rules and Policies, that it conducts research with user data). Therefore, it can be argued that data pertaining to the author of a tweet can be processed for research purposes on several bases: consent, legitimate interest, and (in countries where such legal basis is available for research) public interest. Even sensitive data may be lawfully processed in this context on the ground that such data have been made manifestly public by the data subject (Article 9.2(e) of the GDPR). However, it is recommended to stop processing tweets access to which have been restricted by the user; such action should be interpreted as withdrawal of consent or objection to the processing. Moreover, the processing of tweets still needs to meet all the requirements

of the GDPR (such as security and accountability), especially transparency (unless the applicable national law provides for an exception). As regards the latter, making the information publicly available is a reasonable solution in cases where contacting every user individually would require disproportionate effort (see Article 14.5(b) of the GDPR).

6 Use of Twitter API for research purposes

Albeit it generally seems permissible under Article 3 the DSM Directive, the use of Twitter data for language research purposes is still associated with considerable organizational effort and lack of legal certainty. In this context, simply obtaining specific permission from Twitter may be a reasonable alternative to relying on statutory exceptions. This could clear any copyright-related issues and diminish the burden related to the GDPR -- when the processing is carried out solely through an API provided by Twitter, it can be argued that Twitter is at least a joint controller for the processing.

Twitter has been offering access to APIs for mining tweets for a long time. Recently, in July 2020, Twitter launched a new version (v2) of its API. Reportedly, academic researchers were one of the largest groups of the API users; for this reason, in January 2021, Twitter launched a new Academic Research product track, allowing researchers to get preferential access to the API.

In theory, this Track allows for a 10 000 000 monthly tweet volume cap (compared to 500 000 in the general track). However, this also depends on the streaming endpoint limits, which reportedly are not entirely up to this standard yet (although they are expected to be raised soon). Moreover, it is also possible to use more detailed queries and rules (1024 characters per query/rule in the Academic Research product track, as opposed to 512 in the Standard track). Finally, Twitter Development Agreement allows academic researchers to distribute an unlimited number of Tweet IDs and/or User IDs if they are doing so on behalf of an academic institution and for the sole purpose of non-commercial research (otherwise, 'only' 1 500 000 Tweet IDs per 30-day period can be shared).

The Academic Research product track is available to graduate students, PhD students, post-docs, faculty or research-focused employees at an academic institution or university with a precisely defined research objective and pursuing non-commercial purposes. To apply for the track, a researcher has to answer a very detailed questionnaire including questions about the project, its funding, methodology, the planned use of Twitter data and ways of sharing the outcomes. Arguably, some may see this questionnaire as intrusive and unacceptable from the point of view of academic freedom.

Access to the Track is free of charge. There is no information available as to how many requests are granted. Like anyone with access to the API, successful candidates are bound by the Twitter Development Agreement and Policy. These documents strictly prohibit any attempt to exceed or circumvent access limitations (rate limits). Moreover, Twitter retains the right to immediately terminate or suspend access to the API at any time and for any reason. It can be expected that any attempt to exceed the permissions granted by Twitter, also based on the above-mentioned statutory exception for text and data mining, will be met with the termination of access to the API.

References

Wasim Ahmed. Using Twitter as a data source: an overview of social media research tools (2019). Available at https://blogs.lse.ac.uk/impactofsocialsciences/2019/06/18/using-twitter-as-a-data-source-an-overview-of-social-media-research-tools-2019/ (26.4.2021).

Nicolas Gold. Using Twitter Data in Research. Guidance for Researchers and Ethics Reviewers. University College London (2020). Available at https://www.ucl.ac.uk/data-protection/sites/data-protection/files/using-twitter-research-v1.0.pdf (27.4.2021).

Pawel Kamocki. 2020. When Size Matters. Legal Perspective(s) on N-grams. Proceedings of CLARIN Annual Conference 2020. 05 – 07 October 2020. Virtual Edition. Ed. Costanza Navarretta, Maria Eskevich. CLARIN, 166-169.

Sarah Perez (2017). Twitter officially expands its character count to 280 starting today. Available at https://techcrunch.com/2017/11/07/twitter-officially-expands-its-character-count-to-280-starting-today/?guccounter=1 (26.4.2021).