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

þÿKamocki, PaweB

Linköping University Electronic Press
2022-07-08

Kamocki , P , Hannesschläger , V , Hoorn , E , Kelli , A , Kupietz , M , Linden , K & Puksas , A 2022 , Legal Issues Related to the Use of Twitter Data in Language Research . in M Monachini & M Eskevich (eds) , Selected papers from the CLARIN Annual Conference 2021 . Linköping Electronic Conference Proceedings , vol. 89 , Linköping University Electronic Press , Linköping , pp. 68-75 . https://doi.org/10.3384/ecp1897

http://hdl.handle.net/10138/350015
https://doi.org/10.3384/ecp1897

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. The first part of this paper provides an analysis of these issues, whereas the second part discusses two possible strategies to address them: using the new Academic Research product track, which enables authorized researchers to access Twitter API on a preferential basis, or relying on the new statutory copyright exception for Text and Data Mining for research purposes.

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. With nearly 400 million users worldwide[1] and over 500 million tweets per day[2], Twitter is one of the most popular platforms for academic research on social media data.

The main research methods used on social media and Twitter data are: 1) Content Analysis for systematically labelling text, audio, and visual communication from social media; 2) Thematic Analysis

[1] After: https://backlinko.com/twitter-users (last visit: 4 April 2022).

[2] After: https://www.internetlivestats.com/twitter-statistics/#trend (the number of 500 million tweets per day was already reached in 2013; the same source indicates that as of 4 April 2022 there are on avarage 9945 tweets per second, which would suggest a much higher number of daily tweets – almost 860 million).

Pawel Kamocki, Vanessa Hannesschläger, Esther Hoorn, Aleksei Kelli, Marc Kupietz, Krister Lindén and Andrius Puksas 2022. Legal Issues Related to the Use of Twitter Data in Language Research. *Selected papers from the CLARIN Annual Conference 2021. Ed. by Monica Monachini and Maria Eskevich.* Linköping Electronic Conference Proceedings 189, pp. 68–75. DOI: https://doi.org/10.3384/9789179294441

locating patterns within data through data familiarisation, coding, developing and revising themes; 3) Social Network Analysis to measure and map the relationships between individuals, organisations, and other entities; 4) Machine Learning teaching computers with pre-labelled subsets to code the remainder of the data, 5) Discourse-linguistic analyses of the language and treatment of socio-political and other issues in Twitter and the comparison with other media. 6) Semantic Analysis examining the meaning of and the relationship between occurrences of words, phrases, and clauses and 7) Time Series Analysis for plotting the frequency of items or events in the above across time (for further information, see Ahmed 2019).

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. Moreover, the use of Twitter and its content is subject to the detailed Terms of Service, which also regulate data re-use and sharing. This paper provides a brief analysis of the above-mentioned issues.

2 Legal Issues in Twitter Data

2.1 Copyright in Tweets

A text is protected by copyright if it is original, i.e., if it constitutes the author’s own intellectual creation (CJEU, case Infopaq, C-5/08). Very short texts, such as slogans or titles are often considered unoriginal, because intellectual creation can hardly manifest itself in a very short format. However, the Court of Justice of the European Union ruled that snippets of 11 consecutive words can potentially qualify for copyright protection (idem). This should not be interpreted as a strict measure of originality, but rather as a guideline; on the one hand, not all 11-gramms are protected by copyright, and on the other: some shorter snippets can also qualify for copyright protection. For instance, according to the opinion of advocate general Szpunar (2018) ‘”All quiet on the Western Front”, declared what is probably the most well-known military report in the history of literature. Featured in the novel by Erich Maria Remarque bearing the same name, this phrase naturally enjoyed, together with the work as a whole, copyright protection’. Kamocki (2020) argues that only n-grams that are no longer than 3 words can safely be regarded as copyright-free.

The maximum length of a tweet is currently set at 280 characters (increased from 140 in November 2017), which corresponds to about 50-60 words in English. This is well enough to be protectible by copyright. However, it has been shown that in practice very few tweets reach the maximum length, and most of them are in fact considerably shorter: an average tweet in English has been reported to be only 33 characters long, i.e. approximately 6-7 words (Perez, 2017). Nevertheless, even this shorter length does not allow to exclude average tweets from copyright protection.

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’) certainly fail to meet the originality criterion. However, a pack of several thousand tweets is likely to contain at least some copyright-protected material (even if it does not include photographs or other media). Since in many cases it is impossible to determine whether a tweet is or is not protected by copyright, it is prudent to consider them as being under copyright. Therefore, especially in analysing tweets en masse, copyright issues have to be observed.

This conclusion has two important implications: one related to the moral rights of authors, the other to the economic rights. Concerning moral rights, Article 6bis(1) of the Berne Convention provides authors with ‘the right to claim authorship of the work (a.k.a. paternity right – added by authors) and to object to any distortion, mutilation or other modification of, or other derogatory action in relation to, the said work, which would be prejudicial to his honor or reputation (a.k.a. integrity right – added by authors)’. Even if one sets the integrity right aside (arguably, using a tweet in research is never prejudicial to the author’s ‘honor or reputation’), the paternity right still obliges researchers to mention the name (or nickname) of the author of every tweet whenever it is quoted or otherwise shared. When tweets are used in bulk, this may lead to a phenomenon known as ‘attribution stacking’.

More importantly, the authors have the exclusive right of reproduction and communication to the public of their works. These two economic rights, harmonised in the EU by the 2001 InfoSoc Directive 2001/29/CE (respectively Article 2 and 3), grant the authors of copyright-protected tweets control over their re-use; in other words, such tweets can only be copied and shared if the author grants permission to do so, or if a statutory exception applies. Both hypotheses are discussed in Section 3 of this paper.

2.2 Tweets as Personal Data

Having established that tweets are potentially copyright-protected, it is now time to examine if they should also be regarded as personal data. Personal data is defined as ‘any information related to an identified or identifiable natural person’ (Article 4, (1) of the GDPR). As per WP29 Opinion 4/2007 on the concept of personal data, information ‘relates to’ a person if it is *about* that person (p. 9).

Tweets necessarily contain information about the author: at the very least the user ID, but possibly also location data or other identifying content (e.g. information about the author’s opinions, preferences, etc.). Therefore, they should be regarded as personal data (see e.g., Gold, 2020) and their processing needs to follow the GDPR, even despite the fact that Twitter is an American company (as per its Article 3.2, the GDPR applies to foreign companies which offer services to EU citizens).

Probably the most important implication of this is the fact that the processing of tweets (including their copying, storage, analysis, anonymisation or any form of sharing) needs a legal basis in order to comply with one of the main principles of data processing under the GDPR, namely the principle of lawfulness (Article 5.1 (a) and Article 6 of the GDPR).

Contrary to a common misconception, consent of the data subject (i.e., the person that the data refers to) is not always necessary; it is only one of the available options. Moreover, consent does not have to be given in writing or even (in principle) be explicit, it can also be implied, inferred from an unambiguous affirmative action. Since Twitter provides its users with the possibility to fine-tune their privacy settings, including public availability of their tweets and profile information, mere making tweets publicly available may arguably be interpreted as granting consent to their processing for research purposes, taking into account that Twitter also expressly informs the users (in its Rules and Policies) that it conducts research on data. A problem with this approach arises, however, when the user deletes a tweet, or changes its parameters in such a way that it is no longer public. This should probably be interpreted as withdrawal of consent (under the GDPR, consent can be withdrawn at any time, cf. Article 7(3) of the GDPR). In such a case, the processing of such tweets should stop (see: EDPB Guidelines 05/2020 on consent under Regulation 2016/679, para. 117), and they should be deleted from the corpus, which is a major (at least organisational) obstacle.

As mentioned above, alternative legal bases are also available. One of such alternatives, often regarded as compatible with research purposes, is ‘legitimate interest of the controller’ (Article 6.1(f) of the GDPR). In order to rely on this basis, the controller should perform a ‘balance of interests’ test to assess whether the interests of the data subject do not override the controller’s interest (e.g. in using the data for research purposes). According to the WP29 Opinion 06/2014 (p. 55), the balancing test should take into account such elements as the reasonable expectations of the data subject (cf. also Recital 47 of the GDPR), the nature of the data, and the potential impact of the processing on the data subject. In the context of language research on Twitter data, it seems that the outcome of the balancing test will likely be in favour of the processing, considering that the data in question are short messages made public by the data subject, that the data subject should be aware that they can be used for research purposes, and that language research is highly unlikely to have any negative impact on the data subject. However, also in this case it can be argued that in a situation where the data subject subsequently deletes the tweet or restricts access to it (making it invisible to the general public), the balance is tilted in the opposite direction, and the controller can no longer rely on the ‘legitimate interest’ legal basis to process the tweet. Furthermore, when the processing is based on ‘legitimate interest’, the data subject has the right to object to the processing (Article 21 of the GDPR) – in such a case, the processing can continue only if it passes a stricter test for ‘compelling legitimate grounds’. For now, in the absence of any EDPB guidelines on the right to object, still relatively little is known about this right and the consequences of its exercise.

Yet another legal basis that can be relevant for the processing of Twitter data for language research purposes is ‘public interest’ (Article 6.1(e) of the GDPR). In order to be able to rely on this ground, processing has to be based on an interest clearly laid down in the law (typically, this basis is used e.g. by tax authorities). In some CLARIN countries, such as Finland or Norway, whose national laws contain specific provisions to this effect, this basis is available and recommended for researchers.

Rather exceptionally, tweets may also contain special categories of personal data (the so-called 'sensitive data', cf. Article 9 of the GDPR), i.e. data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, as well as genetic data, biometric data, and data concerning health, sex life or sexual orientation. The processing of such data must be based on specific, strictly defined legal bases (for example, consent for the processing of such data has to be 'explicit', and so our analysis of 'consent' as a legal basis for processing Twitter data for research purposes cannot apply). However, one of these specific legal bases for the processing of sensitive data applies to Twitter data: according to Article 9.2(e), sensitive data can be lawfully processed if they have been 'manifestly made public by the data subject'. Rather obviously, publicly tweeting about e.g., one's health (by announcing an operation or a diagnosis) does count as making this information 'manifestly public', and therefore the above-mentioned legal basis can apply. However, it is much less obvious whether it continues to apply after the relevant tweet has been deleted or restricted. In any case, given the sensitive nature of those special categories of personal data, researchers should always be very prudent while relying on this legal basis.

Tweets may also contain personal data related to third persons, i.e. individuals other than the author of the tweet. Although a lot depends on the circumstances of every specific case, in general it is rather difficult to find an appropriate legal basis for the processing of such data. In particular, it seems difficult to rely on consent (as usually nothing indicates that the third person has consented to her personal data being published in the tweet), or legitimate interest (as the third person has little reason to expect that her data, tweeted by someone else, will be used for research). Unfortunately, it is rather impossible to automatically detect tweets containing third persons' personal data, which further complicates the use of Twitter data for language research purposes.

Even if the processing complies with the principle of lawfulness (i.e., it has an appropriate legal basis), there is a number of other requirements in the GDPR that it has to meet. One of such requirements is related to the principle of transparency, under which the data subjects should be provided with information about the processing in a concise, transparent, intelligible and easily accessible form (cf. Articles 12 and 14 of the GDPR). This information includes, inter alia, the identity and contact details of the data controller, the purposes for which the data are being processed, the data retention period, and the rights of the data subject (for more information, see WP29 Guidelines on transparency (WP260rev01)). In the context of Twitter data analysis, taking into account the sheer amount of processed data and concerned data subjects, this principle seems particularly difficult to observe. However, the GDPR, in its Article 14.5(b), includes an exception from this principle for cases where provision of the information proves impossible or would involve disproportionate effort. As per the Article 14.5(b) itself, this exception can apply in particular to the processing carried out for research purposes. In assessing whether the necessary effort is disproportionate, according to the Recital 62 of the GDPR, account should be taken of such elements as the number of data subjects (the higher the number, the bigger the effort), the age of the data (the older the data, the bigger the effort) and any appropriate safeguards adopted by the controller (e.g. pseudonymisation, encryption, restricted access to the collected data, etc.). This assessment has to be made on a case-by-case basis, but it seems that when analysing Twitter data for language research purposes at least the first element – the number of data subjects – will generally weigh in favour of the exception of 'disproportionate effort'. It should be noted that even if the exception applies, the controller should still 'take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available'. A reasonable solution would be to publish the relevant information e.g. on the project's website.

2.3 Tweets and Contracts

In order to be able to tweet, one needs to create a Twitter account and accept (among other documents, such as the Privacy Policy) Twitter's Terms of Service (ToS)[3]. Upon acceptance, the ToS become a binding contract that both the user and the platform provider are bound to respect.

It has been demonstrated *supra* that tweets can be protected by copyright. Therefore, it is particularly interesting for further analysis to examine how copyright issues are addressed in the ToS. In Section 3

[3] Available at https://twitter.com/en/tos#intlTerms (last visit: 9 February 2022).

of the ToS, the paragraph entitled ‘Your Rights and Grant of Rights in the Content’ provides the following:

‘By submitting, posting or displaying Content on or through the Services, [the user] grant[s] [Twitter] a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods now known or later developed (for clarity, these rights include, for example, curating, transforming, and translating). This license authorizes [Twitter] to make [the user’s] Content available to the rest of the world and to let others do the same’.

This means that although the user retains copyright in his tweets, he grants Twitter a very broad permission (license) to re-use them for free on a non-exclusive basis (i.e., the user can still use the tweets himself and authorise others to do so). As a consequence, someone who would like to copy and share tweets can receive the necessary authorisation either directly from the user (which in most cases is unworkable in practice, given the sheer number of Twitter users) or from Twitter (a license from Twitter would be a *sublicense*; sublicensing is explicitly authorised by the ToS). Also, theoretically, nothing prevents the users from re-publishing their own tweets outside of Twitter, including e.g. in .xml format, and under an open license.

Twitter ToS also grants every user access to the Twitter Services, but no general sub-license to re-use the content (a limited personal license is provided only to use the software provided as part of the Services, for the sole purpose of enabling the user to enjoy the services). Moreover, certain uses and actions are expressly forbidden: for example, the user is not allowed to ‘access or search or attempt to access or search the Services by any means (automated or otherwise) other than through (...) currently available, published interfaces that are provided by Twitter (and only pursuant to the applicable terms and conditions)’. Interestingly, it is allowed to crawl the Services (i.e., presumably, to use Twitter to find, discover and visit URLs) in accordance with the provisions of the robot.txt file (such uses can hardly be efficiently prohibited); it is not allowed, however, to scrape the Services (i.e. to extract the data) without Twitter’s prior permission. It seems therefore that mining of tweets without specific permission, 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, or perhaps even to a lifetime ban. Theoretically, Twitter could also sue for damages for breach of contract, but this, in our opinion, is highly unlikely to happen, for at least two reasons 1) the economic loss suffered by Twitter would probably be negligible, if even possible to quantify, and therefore the amount of compensatory damages that could be claimed by Twitter would also be negligible; 2) suing a non-commercial research organisation such as a university would result in controversies that may be harmful to the image of the company. Despite its low probability, the fear of legal action from Twitter is probably the reason why those researchers who have indeed scraped data from Twitter are not transparent about it, e.g. in their ethical self-assessments, and therefore many Twitter corpora remain underexplored and ‘under the radar’.

Although the authors have not tested this in practice, they assume that scraping tweets is not only forbidden by Twitter ToS, but also made impossible (or at least very difficult) by technological protection measures (TPM). For example, Twitter may detect ‘unhuman’ use of its Services (such as consulting a very large amount of URLs in a very short period of time from one IP address) and prevent it. Circumventing such technological measures is not only expressly prohibited by the Twitter ToS, but also in principle forbidden by law (cf. Article 6 of the InfoSoc Directive).

Another provision of the ToS that is noteworthy from the point of view of this analysis is related to the termination of the contract. In Section 4, the paragraph entitled ‘Ending these Terms’ provides that Twitter may suspend or terminate the user’s account or cease providing the user with all or part of the Services at any time ‘for any or no reason’ [sic!]. Therefore, there is no legal guarantee that Twitter will continue to provide its Services in the future, which is a major problem from the point of view of sustainability of Twitter data used for research purposes.

3 Using Twitter Data for Language Research – Possible Strategies

In the previous section, it has been demonstrated that in order to be able to lawfully scrape and analyse tweets for language research purposes, researchers need, in addition to observing GDPR principles, to obtain a specific permission from Twitter, or alternatively to rely on a statutory exception. Both solutions have their advantages and disadvantages, as discussed in this section of the paper.

3.1 Twitter API for Academic Research

The use of Twitter data for language research purposes is still associated with considerable organisational effort and lack of legal certainty. In this context, simply applying for a specific permission from Twitter may be a reasonable solution. This could not only clear any copyright-related issues, but also 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 a joint controller for the processing.

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 has launched a new Academic Research product track, allowing for a preferential access to the API (Tornes and Trujillo, 2021).

In theory, the Aademic Research track allows for a 10 000 000 monthly tweet volume cap (compared to 500 000 in the general track), although 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). In addition, the 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 content itself, however, cannot be shared. This might be seen as an inconvenience, but it allows to solve many GDPR-related problems with tweets that have been deleted or restricted by the user (see above).

The Academic Research product track is available to: 1) researchers, post-docs, professors of fellows at academic institutions (undergraduates are expressly excluded); 2) Master's students working on theses; 3) PhD candidates working on dissertations and 4) persons affiliated with an academic institutions and working on a clearly defined research project. In all cases, the applicant has to pursue a non-commercial purpose, and have a Twitter account.

In the process of applying for the Academic Research track, the applicant has to prove his or her affiliation with an academic institution (by providing a link to the webpage on his or her institution's website listing his or her name, or to his or her Google Scholar profile), provide information about the institution, his or her department or lab, and his or her current role in the research group. Then, the applicant is asked to answer a very detailed questionnaire about his project including questions about its name, 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. There is no information available as to how many requests are granted, and what are the admission criteria. Successful candidates are bound, like anyone with access to the API, 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 will be met with termination of access to the API. This, combined with the possibility for Twitter to modify or stop providing its Service at any time, is far from optimal from the point of view of sustainability of research data accessed via the API.

3.2 Statutory Exception for Text and Data Mining for Scientific Research Purposes

As explained above, re-use of copyright-protected content is only possible if it is authorised by the author (directly or indirectly), or if it is exempted from authorisation by a statutory exception. Recently the Directive 2019/790 on copyright in the Digital Single Market (DSM Directive) introduced (in its Article 3) a new statutory exception, supposed to cover such scenarios as using Twitter data for language research purposes at research institutions. The Directive is now transposed in most EU Member States (the deadline for transposition was set for 7 June 2021, but in many countries the relevant legislative processes were delayed by the COVID-19 pandemic).

This new exception allows research organisations (such as universities) and cultural heritage institutions (such as libraries, museums or archives) to make copies of copyright-protected content in order to carry out text and data mining for scientific research purposes (including in public-private partnerships). The exception only applies to content to which the above-mentioned institutions have ‘lawful access’. This requirement is often presented as a hurdle, but when it comes to publicly available tweets, the criterion is easily met: as per Recital 14 of the DSM Directive, ‘Lawful access should also cover access to content that is freely available online’.

In general, copyright exceptions are overridden by contracts. In other words, if a contract (such as Terms of Service) prohibits certain uses (like scraping), this prohibition remains in principle unaffected by statutory exceptions. However, the exception for text and data mining for research purposes has a rare and very important feature: it is not overridable by contracts, i.e. any contractual provision contrary to this exception is unenforceable (Article 7.1 of the DSM Directive). This means that the beneficiaries of the exception (research organisations and cultural heritage institutions) may scrape Twitter data, despite the general prohibition of scraping in the Twitter ToS. It remains to be seen if such use will be tolerated by Twitter which, as per the ToS, can cease to provide the Services to any user for any reason, including for no reason at all (see above). In this context, the exception can shield against a legal action from Twitter for breach of contract, but not against unilateral termination of the contract by Twitter.

Another aspect of the exception concerns its relation with technological protection measures. This seems to be the biggest grey area of the exception, as Article 3.3 of the DSM Directive allows platform providers to apply technological measures to disable text and data mining for research purposes, but only to the extent necessary to ensure the security and integrity of their networks and databases. In our opinion, Twitter would have a good chance to succeed in arguing that TPMs implemented to prevent unauthorised scraping are, in fact, necessary to achieve such goals, as unlimited scraping might place too heavy a burden on their servers and affect the accessibility of their services for other users; however, it still remains to be seen how this issue will be worked out in practice. According to the DSM Directive, Member States shall encourage stakeholders to define commonly agreed best practices in this area.

The copies made under the exception (i.e., corpora) have to be stored ‘with the appropriate level of security’ to protect them against unauthorised access. They can, however, be re-used in other projects or for evaluation purposes. Unfortunately, the exception itself 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 (Section 60d of the German Copyright Act) allows for the corpus to be shared with a limited circle of persons for joint scientific research, which seems to allow sharing within research infrastructures such as CLARIN.

The advantage of relying on this exception rather than the Twitter API in using Twitter data for language research purposes is a greater degree of autonomy and control over the data collection process. On the other hand, it remains an uncharted territory with many great areas. Unlike the use of Twitter APIs, the statutory exception also does not provide any relief regarding the GDPR compliance.

4 Conclusion

Twitter data present a number of legal issues: tweets can be protected by copyright, they contain personal data, and access to them is regulated by the Terms of Service. This, however, does not mean that Twitter data are out of reach for language researchers. Quite the contrary, there are at least two ways to get hold of such data: via the API provided by Twitter (with a specific, preferential track dedicated to academic research), or by relying on the new copyright exception for Text and Data Mining. None of these approaches is fully satisfactory. Moreover, taking into account the specificities of national laws (especially with regards to the Text and Data Mining exception), identified research questions and adopted research methods, specific solutions for handling Twitter data should still be adopted on a case-by-case basis.

One such solution that looks quite tempting might be a hybrid approach between the Twitter API and the statutory exception; in this scenario, the data are accessed via the Twitter API, then copied on the basis of the copyright exception for Text and Data Mining, anonymised, stored, re-used and potentially even shared with other researchers (this last element seem to depend mostly on the applicable national law). As both the Academic Research track in the Twitter API and the relevant copyright exception were only introduced relatively recently, best practices have yet to emerge – also by trial and, quite inevitably, by error.

References

Ahmed, W. 2019. 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/ (09.02.2022).

Gold, N. 2020. *Using Twitter Data in Research. Guidance for Researchers and Ethics Reviewers*. University College London. Available at https://www.ucl.ac.uk/data-protection/sites/data-protection/files/using-twitter-research-v1.0.pdf (09.02.2022).

Kamocki, P. 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. Available at https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf (09.02.2022).

Perez, S. 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 (09.02.2022).

Szpunar, M. 2018. Opinion of Advocate General Szpunar delivered on 25 October 2018. Case C-469/17. Funke Medien NRW GmbH v Bundesrepublik Deutschland. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1574794094419&uri=CELEX:62017CC0469#t-ECR_62017CC0469_EN_01-E0002 (09.02.2022).

Tornes, A., Trujillo, L. 2021. Enabling the future of academic research with the Twitter API. Available at https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api.html (09.02.2022).