Text and Data Mining in the context of Smart Data – an economic perspective

Patrick Bunk, Ubermetrics Technologies GmbH

Data value chains are a necessity to provide equal access to key technologies of the digitization process for small and medium-sized companies and enable them to take part in the European data economy that is about to begin. This perspective is not supported by the current position of the Federal Government on the European copyright reform and the German regulation on text and data mining (TDM).

It is currently being discussed at European level whether a newly created special permit for TDM should be restricted to research facilities purely for the purpose of research.[i] As a consequence, it is expected that all TDM processes that happen every day on everyone's computers and smartphones will be considered copyright violations unless the user acquires additional TDM-licenses for all the material he or she already has legal access to.[ii]

At the same time, this protection is planned to be extended to snippets, small text excerpts from products of press publishers, without the pre-condition of artistic expression that is usually required for copyright to apply.[iii] This would protect all ordinary texts, including all arbitrarily short sequences of words, as long as they are published in a press publication. These then would be withdrawn from public use for 20 years.

In Germany, the so-called "Urheberrechts-Wissensgesellschafts-Gesetz"[iv] also covers the inclusion of a TDM permit in § 60d UrhWissG-E for non-commercial research purposes only. The increasing use of TDM technologies by companies and start-ups could thus be complicated by the rather impractical requirement of obtaining licenses and thereby diminishing competitiveness on international markets.

In order to understand the effects of these reform projects on data value chains, it is helpful to first examine the definition of text and data mining.

1. Text and Data Mining

Text and Data Mining (TDM) is initially, by definition, any process that extracts high-quality information or connections from texts or data. These processes have been an integral part of computer science for decades and can be found in countless fields of application.

A classic example of TDM is the search function on a Windows, Mac or Linux operating system. This function analyses all documents on the PC, makes copies of all sentences in all documents and saves them in a structured database. As soon as the user searches for a document by entering any part of the document into a search mask, the information is retrieved from the database in fractions of a second, not from the original document.

There are numerous other examples of TDM in the present time:

- Spelling and grammar correction and many other machine analyses of human language as well as any machine translation of texts
- Pattern recognition methods, e.g. any that makes it possible to click on a phone number in e-mails on the smartphone
- Trend recognition and analysis
- Spam detection
- Internet search engines such as Bing, Google, Qwant or Cliqz

Most artificial intelligence (AI) methods are by definition TDM technologies, because they learn higher-quality information as rules from the sets of data or text.

Recent breakthroughs over the last 5 years have shown that AI technology is capable of performing monotonous information extraction and classification tasks. Thus, these

technologies can for example sort texts according to the language they are written in, identify an account number on an invoice, or animals in images.

2. How do today's artificial intelligence algorithms work based on deep learning?

The technology creates a simulated version of a neuron, a human brain cell, and simulates it. In practice, hundreds to millions of these neurons are created. The AI technology shows this network of interconnected neurons certain input data, for example animal images or sentences, together with a classification from which the algorithm is then supposed to derive rules. For this purpose, one has to show these algorithms different classified objects until some of the simulated neurons learn some aspects of the problem. Once this is achieved, this algorithm can perform tasks that previously could only be performed by humans.

However, there is a prerequisite: you need a very large amount of data as the starting point to allow the learning of patterns. These are usually hundreds of millions to billions of texts and images. SMEs in particular do not have such large, specifically created data sets. But these are easily available for very few large IT companies like Google and Facebook. All other users in science and business therefore use the largest data set in the world – the open Internet including all data sources that provide free access – as a basis on which the AI technologies can learn structures in the data, for example the Wikipedia.

3. The current legal situation

At present, there is still controversy as to whether a copyright license has to be obtained for TDM, as also the government draft of the UrhWissG-E confirms: *"the automated evaluation itself, the core of the so-called text and data mining, is not a copyright-relevant act".*

This statement, however, does not reflect the practice of TDM in real life. In order to develop, evaluate or improve TDM methods, a consistently large body of data is necessary to be able to measure the quality of the algorithm. This pre-structured corpus always consists of a large number of corrected documents or data series representing the problem. These are at least temporary copies, which naturally fall under the copyright law. Theoretically, unauthorized use is possible for brief, incidental copies when they are required as an integral and essential part of a technical process, whose sole purpose is a legitimate use and these copies have no independent economic significance (cf. Section 44a German Copyright Act). This exception can so far only be used for those TDM methods whose corpus is deleted immediately after the information extraction.[v] However, this would prevent any new development, evaluation or improvement of TDM algorithms in Europe. Moreover, in the field of application of TDM algorithms, it is frequently necessary from a user perspective or from data protection principles to be able to compile the obtained analytical results in a comprehensible manner by disclosing the original sources to validate the algorithms work.[vi]

The problem presents two key challenges for TDM technology-using companies:

1. They must determine whether millions of texts are copyrighted. For this purpose, an individual evaluation is necessary in order to assess whether the respective text constitutes a individual creative creation, with the protection being relevant not only for creations of peculiar character but also for works of little creative value ("small coin"). [vii]
2. In order to obtain usage licenses, they must determine who the authors or rights holders of the exploitation rights are.

Considering the fact that the goal of the TDM deployment is generally simply the collection of information - a process which is freely

carried out by humans - the question arises as to whether copyright protection is appropriate here. The purpose of copyright protection is the protection of intellectual creation and not of information.[viii] Information as such should remain public in the telos of the existing protection system so as not to hinder social exchange and progress.[ix] As long as the original work utilization is not affected by the TDM, a TDM barrier would only have to ensure that the TDM does not become the gateway for other types of copyright-relevant usage.[x]

The draft of the EU commission states: *"Text and data mining may also be performed with respect to mere facts or data that are not protected by copyright, and in such cases no authorization would be required."*[xi] Unfortunately this implies, according to the German Federal Second Chamber, that in all cases where there exists a risk that some text or data involved may be copyright protected, one must obtain prior consent before applying TDM from the rights holders of these text snippets.

4. Liability risk reduction by filtering protected works as an alternative approach?

As a result of this regulatory proposal, companies will have to ask themselves whether it is possible to avoid the liability risks by pre-filtering all potentially copyrighted works. Unfortunately this is economically infeasible: Even in an optimistic estimate one would assume that 1% of 1% of all documents in a corpus could contain copyright-relevant parts, then anytime a TDM process runs on this corpus this would constitute a copyright infringement of these documents. With corpus sizes of 10 billion texts one should expect that one will have to infringe in 1 million cases. This implies a liability risk of several million Euros per TDM process for common commercial claims. With such huge economic risks, TDM processes in Europe will not even be allowed in research departments of companies.

Therefore, the proposed EU regulation would require the establishment of a comprehensive copyright filtering infrastructure. However, this is currently not possible.

In order to determine whether a text contains a sufficiently large copy of another copyrighted text of a third party, one would have to store all the texts of all copyright holders in order to compare them with the text in question - this process itself is already a copyright infringement according to the current interpretation of European law.

Optimists may argue that computer science would surely find a solution. The absurdity of the regulatory proposal, however, is that even in the event that this would be algorithmically possible, the examination of whether an article is copyrighted is by itself a TDM process. This means that the examination process itself, in any case where it would identify a copyrighted part in a document, would have caused a copyright infringement through the analysis of the copyrighted text.

Consequently, it must be concluded that the filtering of copyright material is not an option for risk minimization. In practical terms, this means barriers for the competitive development of TDM and AI technologies in Europe. An aspect that would be diametrically opposed to the development of a European data economy postulated by the European Commission.

5. Competitive impact

A further consequence of this proposal is a permanent change in the structure of competition in the field of AI technologies.

As noted, it is impossible to prove that the training sets for AI technologies are free of copyrighted material. Therefore, licenses are required for very large data sets in order to be able to develop AI technologies. For companies without a dominant market position, e.g. SMEs and start-ups, it is virtually impossible to negotiate contracts with every copyright owner in Europe. An exception for start-ups does not solve this problem, since even among the very successful start-ups, almost none of them has a dominant market position. This

means for all other SMEs and start-ups that the transaction costs for the negotiation with every copyright holder in Europe are prohibitively high and that the asymmetry of power in the negotiation case also makes a market solution socially inefficient.

What happens to the big US players like Google?

Google crawls the Internet and trains its algorithms under the US Fair-Use Doctrine[xii] in the USA. The Fair-Use doctrine allows the use of copyrighted material, taking into account the following criteria:

- Purpose and nature of the use, including whether the use is of a commercial nature or for non-profit educational purposes;
- the nature of the protected work;
- extent and materiality of the used part in relation to the total work, and
- the effect that the use in the potential market can have on the value of the protected work.

The EU Commission proposal does not have any effect in this regard. Even a legal "rectification", if at all possible, would not change anything because of the economic constraints of the authors.

Google is a very large search engine with many users in Europe. To be found on the internet is still very important for all companies and authors, especially for publishing houses. Therefore, standard contracts are a viable option for companies that are already dominant in the market.[xiii] Because of the link to the search function, it is obvious that there is a high willingness to sign this standard contract. Smaller companies, including start-ups in particular, will have to negotiate with each right holder individually and not get as favourable terms. In the area of future technologies such as AI, monopolies will therefore be strengthened and market entry barriers created for European companies without any anti-competitive behaviour of the market-dominant companies.

The German ancillary copyright law for publishers is a real life example of the this economic absurdity.

In this case, Google received free licenses from all German publishers to display snippets (small text excerpts). The competition authorities did not see any infringement of competition law.[xiv] At the same time, many publishers demanded licenses from all smaller providers.[xv] As a result, small suppliers were disadvantaged and the competition in this market is shut down for the long term.

In addition, publishers could shift competition from the primary to the secondary market of the data processors and information intermediaries. If, for example, search engine providers, such as Cliqz, which is financed by Burda Verlag, do not receive licenses from the publishing competitor Axel Springer and Qwant cannot access texts from the Burda publishing house, the analysis results can only represent a part of the reality of life and lose quality. If these providers are not able to assert themselves in the market, the market concentration will be on a few market-dominating providers in the long-term and therefore lead to a supply reduction. Securing media and opinion diversity in the search engine sector is, however, an important concern.[xvi]

The establishment of licensing models in which successful TDM or AI technology-using or -developing companies pay part of their profits to copyright owners is only feasible, provided that a profitable business model remains for these companies in the international context as well. However, in order to remain competitive, companies using TDM or AI technologies could shift their place of business to countries with a legal situation comparable to the fair-use doctrine. These profits as well as value-adding jobs for AI developers would not be created in Europe.

As a result, the European economy will depend on AI systems from non-European suppliers. The EU Commission's proposal therefore forces a entrenchment of the natural monopoly

position of an existing non-European market participants, and undermines any chance of European companies to compete in the field of Artificial Intelligence.

6. How can this problem be solved?

Some draft opinions from the parliamentary committees ask for a broader TDM barrier[xvii], but they need to be championed by someone in the trilogue process between the Commission, the Council and the Parliament.

A simple TDM barrier could be based on the fair-use doctrine, compatible with international copyright treaties, while preserving the interests of the authors:

„Any use of the TDM is permitted without the consent of the copyright holder both for research facilities and for private providers and for non-profit and commercial purposes, under the condition that:

- *the access to the original sources is legal or publicly available,*
- *the utilization of the original source is not made more difficult by the analysis as well as the dissemination of the analysis results, in particular no substituting effect occurs,*
- *the work is only used for the extraction of information or any other use authorized by the author. "*

These restrictions could be used to ensure that the respective authors are not affected by the use of TDM or the dissemination of the analysis results in the utilisation of their work.

7. Conclusion

On the way to a European data economy, public information should continue to be accessible to everyone. In addition to very limited exceptions, copyright should continue to protect only the artistic expression, not the facts and factual connections processed in the artistic work.

The proposed TDM regulation kills any chance of an European data economy that is driven by European companies. In the case of legal access to a public document, there should be no difference as to whether the information contained therein is processed by a person or by a machine. Computers and algorithms have no joy in the artistic expression of the work that is protected by copyright law. (At least not yet.)

In the attempt to subsidize only the business models of European publishers with new copyright information on search engine technologies, the EU will instead strengthen the dominance of a few US technology companies and thus increase their negotiating power with publishers. This will, as shown, come at the expense of the competitiveness and the innovative power of European companies.

As a consequence of these plans, existing European competences in key technologies such as data analysis and artificial intelligence systems would be used to the economical benefit of a few US-American technology companies only.

The need to finance journalism is comprehensible. The problem is caused by economic factors. The proposed regulation, however, will reach the opposite of its main objective with regard to the development conditions of the European data economy. Classic publishing business models will be damaged in the medium term and only a handful of market-oriented internet actors will benefit.

[i] Art. 3 of the proposal of the EU Commission: Proposal for a Directive of the Parliament and the Council on copyright in the Digital Single Market, COM(2016) 593 final, vom 14.9.2016.

[ii] Bundesrat Drucksache 535/16, p. 7.

[iii] Art. 11 COM(2016) 593 final.

[iv] Draft law on the approximation of copyright law with the current requirements of the knowledge society (Urheberrechts-Wissensgesellschafts-Gesetz – UrhWissG)

[v] See for the individual requirements: *Triaille/ de Meeûs d'Argenteuil/ de Francquen*, Study on the legal framework of text and data mining (TDM), funded by the European Commission, March 2014.

[vi] Bundesrat Drucksache 535/16, p. 7.

[vii] Bundestag Drucksache. 1V/270, 38.

[viii] *Raue* GRUR 2017, 11 (13).

[ix] *Raue* GRUR 2017, 11 (13).

[x] *Schack* ZUM 2016, 266 (269).

[xi] Draft law of the Federal Government on the approximation of copyright law with the current requirements of the knowledge society, p. 44.

[xii] See for example the legal case *Authors Guild v. Google, Inc.* Court of Appeals for the Second Circuit New York, decision of 16 October 2015 – 13-4829-cv.

[xiii] The Federal Cartel Office did not regard the request for consent to the free display of text excerpts on the part of the press publisher operating on the Internet site as an abuse of the dominant position of a search engine operator, BKartA Bonn, decision of 8 September 2015 – B 6 - 126/14.

[xiv] BKartA Bonn, decision of 8 September 2015 – B 6 - 126/14.

[xv] OLG München, decision of 14 Juli 2016 – 29 U 953/16.

[xvi] Paal ZRP 2015, 34.

[xvii] *Committee on Industry, Research and Energy*, Draft Opinion, of 2 March 2017 available from: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2f%2fEP%2f%2fNONSGML%2bCOMPARL%2bPE-592.363%2b01%2bDOC%2bPDF%2bV0%2f%2fEN; *Committee on the Internal Market and Consumer Protection,* Draft Opinion of 20 February 2017, available from: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2f%2fEP%2f%2fNONSGML%2bCOMPARL%2bPE-599.682%2b01%2bDOC%2bPDF%2bV0%2f%2fEN