Available online at www.sciencedirect.com
ScienceDirect
www.compseconline.com/publications/prodclaw.htm
Computer Law & Security Review

Legal aspects of text mining

Maarten Truyens [a,1,*], Patrick Van Eecke [a,b,c,1,*]

[a] University of Antwerp, Belgium
[b] King's College, London
[c] Institute of Computer and Communications Law, Queen Mary University of London, UK

Keywords:
Copyright
Text mining
Data mining
Reproduction right
Databases

ABSTRACT

"Text mining" covers a range of techniques that allow software to extract information from text documents. It is not a new technology, but it has recently received spotlight attention due to the emergence of Big Data. The applications of text mining are very diverse and span multiple disciplines, ranging from biomedicine to legal, business intelligence and security. From a legal perspective, text mining touches upon several areas of law, including contract law, copyright law and database law. This contribution discusses the legal issues encountered during the assembly of texts into so-called "corpora", as well as the use of such corpora.

1. Introduction

1.1. *What is text mining?*

Text mining or *text analytics* is an umbrella term describing a range of techniques that seek to extract useful information from document collections through the identification and exploration of interesting patterns in the unstructured textual data of various types of documents[2] – such as books, web pages, emails, reports or product descriptions. A more formal definition restricts text mining to the creation of new, non-obvious information (such as patterns, trends or relationships) from a collection of textual documents.[3]

Typical text mining tasks include activities of search engines, such as assigning texts to one or more categories (*text categorisation*), grouping similar texts together (*text clustering*), finding the subject of discussions (*concept/entity extraction*), finding the tone of a text (*sentiment analysis*), summarising documents, and learning relations between entities described in a text (*entity relation modelling*).

Given the sheer amount of data produced each day, text mining applications are on the rise, and can be considered part of the recent "Big Data" and data mining trend. Text mining is related to, but nevertheless different from the better-known subject of *data mining*. Text mining derives much of its inspiration and direction from research on data

* *Corresponding authors*. University of Antwerp, Belgium.
E-mail addresses: maarten.truyens@uantwerpen.be (M. Truyens), patrick.vaneecke@uantwerpen.be, Patrick.VanEecke@dlapiper.com (P. Van Eecke).

[1] Maarten Truyens is researcher at the University of Antwerp. Stadscampus, S.V.121Venusstraat 23 2000 Antwerpen, Belgium. Prof. dr. Patrick Van Eecke is professor of law at the University of Antwerp, visiting lecturer in IT Law, King's College, London, and visiting professor, Institute of Computer and Communications Law, Queen Mary University of London. He is also a member of the CLSR Editorial Board.

[2] R. FELDMAN and J. SANGER, *The text mining handbook: advanced approaches in analyzing unstructured data*, New York, Cambridge University Press, 2007, 1.

[3] M. HEARST, University of Maryland, *Untangling Text Data Mining*, Proceedings of the ACL '99, 1999.

0267-3649/$ – see front matter © 2014 Maarten Truyens & Patrick Van Eecke. Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.clsr.2014.01.009

mining,[4] and shares many high-level architectural similarities.[5] In fact, data mining is not the only neighbouring practice area for text mining – other practice areas include general statistics, machine learning, database management, artificial intelligence and computational linguistics.[6]

In a corporate context, text mining can provide for efficiency improvements and applications that cannot be reached by traditional database techniques, because it is assumed that the majority of a typical company's information is stored in the form of unstructured text, with information scattered over many computers instead of being stored in organized databases.[7] Practical business applications can be thus be very diverse[8]: the automated processing of open-ended questions in electronic surveys; providing automated feedback to customer questions submitted online; compiling frequently asked questions from customers; automated screening of CVs from potential hires; monitoring what customers say about a company's product on social networks; checking large patent databases to avoid patent infringements; and so on.

Text mining thrives in linguistic applications, such as statistical machine translation. For example, on the basis of parallel texts (such as those with EU legislation, EU parliamentary works, international treaties, etc.) carefully translated by humans into various languages, software can learn how expressions are translated in each language.[9] Similarly, software may be able to detect in real-time cyber-bullying during chat conversations, through deep text learning, based on sentiment analysis and text pattern discovery.[10]

Text mining also has many applications in the field of legal research, such as discovery procedures in trials, automatic summarization or argumentation extraction of court decisions,[11] knowledge extraction from legal statutes,[12] and assisting with contract drafting (document automation).[13]

However, the area where text mining currently seems most promising is biomedicine. There, text mining is used to discover previously hidden relationships in existing knowledge. For example, a specific gene may be mentioned briefly in some research articles, but may not stand out on its own. Among thousands of research articles, however, text mining may be able to filter out such gene, and hint that it could provide for an interesting avenue for further research.[14]

1.2. *Corpora*

Text mining spans a range of diverse activities, with equally diverse accompanying workflows. What is common to most types of text mining activities – and is highly relevant from an intellectual property point of view – is the involvement of a *corpus*, which forms the basis of the computations performed by the text mining software.

A corpus can either be created from scratch (*e.g.*, by collecting texts from public web pages, such as those from the EU parliament's website containing preparatory documents in various translations), or can be provided by a third party (typically through license agreements[15]). Many other corpora exist, in varying sizes, domains of application, quality and cost. For example, for linguistic applications, a historically well-known corpus is the Penn TreeBank, a 4.5-million-word corpus that contains texts from four sources, including text samples from a broad range of contemporary American English in 1961, as well as newspaper articles from the Wall Street Journal.[16] Biomedical text mining applications, on the other hand, frequently rely on corpora consisting of medical reports in databases such as PubMed.

While a corpus may simply consist of a collection of plain text documents, it may also include *annotations*, a kind of metadata ("tags")[17] to enrich the text with interpretative linguistic information to help the software "learn" from examples provided by a human domain expert (*e.g.*, a linguist for a machine translation application, or a lawyer for a case law automation tool). Such workflow is called *supervised learning* or *semi-supervised learning*, depending on whether all or parts of the examples were provided by humans.[18]

For example, a corpus editor can decide to apply part-of-speech data to words or word groups in the corpus in order to indicate their lexical function (noun, pronoun, verb, adverb,

[4] For example, both text and data mining rely on pre-processing routines, pattern-discovery algorithms, and visualization tools to interact with the end-user.

[5] R. FELDMAN, o.c., 1.

[6] G. MINER, *Practical text mining and statistical analysis for non-structured text data applications*, Waltham, MA, Academic Press, 2012, 31.

[7] M. KONCHADY, *Text mining application programming*, Boston, Mass., Charles River Media, 2006, 3.

[8] For a general overview, see footnote 7, 11–20.

[9] P. KOEHN, *Statistical machine translation: textbook*, Cambridge, Univ. Pr., 2007, 5–6.

[10] See for example the "AMiCA" project (www.amicaproject.be).

[11] R. MOCHALES-PALAU and M.-F. MOENS, "Argumentation mining," *Artificial Intelligence and Law* 19, 1; A. WYNER, R. MOCHALES-PALAU, M.-F. MOENS and D. MILWARD, "Approaches to Text Mining Arguments from Legal Cases," in E. FRANCESCONI, S. MONTEMAGNI, W. PETERS and D. TISCORNIA (eds.), *Semantic Processing of Legal Texts*, Springer Berlin Heidelberg, January 1, 2010.

[12] C. VANDA BURNS, "Online legal services — a revolution that failed?" October 2007.

[13] *Ibid.*

[14] Example cited by a user submission in I. HARGREAVES, Supporting Document T to the Hargreaves report – Text Mining and Data Analytics in Call for Evidence Responses, May 2011, www.ipo.gov.uk/ipreview-doc-t.pdf, 5.

[15] For linguistic corpora, see for example the Linguistic Data Consortium www.ldc.upenn.edu/.

[16] J. PUSTEJOVSKY and A. STUBBS, *Natural Language Annotation for Machine Learning*, O'Reilly Media Inc., October 10, 2012, 8.

[17] G. LEECH, *Corpus annotation: Linguistic information from computer text corpora*, 1997, 1–18.

[18] Supervised learning typically provides higher quality output, but may have a significantly higher cost due to the amount of human effort required. Practice has shown however that, when certain assumptions can be made about the set of human-provided examples (*e.g.*, regarding their smooth distribution), using unlabelled data in conjunction with a small amount of labelled data can produce very decent results. Semi-supervised learning will be most useful whenever obtaining data points is cheap, but obtaining the human-provided labels costs a lot of time, effort, or money. This is the case in many application areas of machine learning (*e.g.*, billions of web pages are readily available, but to classify them reliably, humans have to read them). See O. CHAPELLE, B. SCHÖLKOPF and A. ZIEN, *Semi-supervised learning*, Cambridge, Mass., MIT Press, 2006, 1–11.