DEval
GERMAN INSTITUTE FOR DEVELOPMENT EVALUATION

DEval Policy Brief 1/2019

TEXT MINING IN EVALUATION

Summary

The global indicator framework to monitor the 2030 Agenda for Sustainable Development pushes for an increase in structured data (e.g. numbers in tables). Yet, the majority of information on development cooperation exists as unstructured text, because processes and impacts are hard to measure by numbers alone.

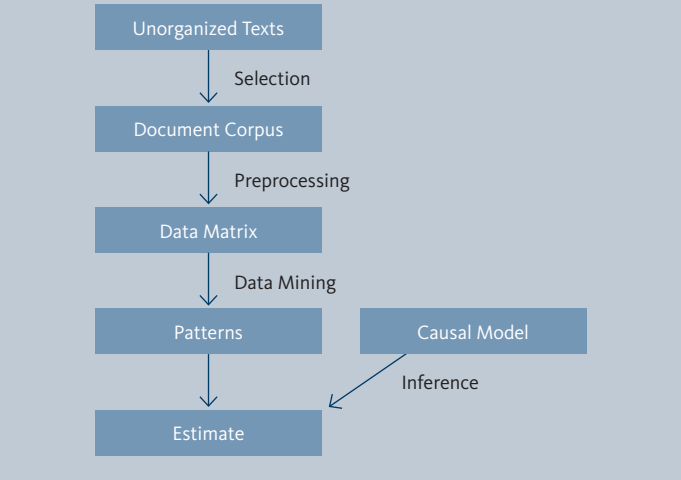
All processes of formulating and implementing development cooperation policies create text: proposals detail means and ends of interventions, progress reports monitor the steps toward the desired outcomes, and evaluations provide evidence on the effects of development interventions. In addition, press and social media reflect the public opinion on development cooperation and, in some cases, are even indicative of outcomes.

The increased availability of digital text has rendered detailed reading of all relevant sources impossible. Additionally, interesting patterns in large collections of documents such as co-occurring terms are often not evident for human readers. Hence, the method of close reading taken in isolation reaches its limits when faced with large bodies of text.

Text mining provides a methodological tool to cope with the abundance of digital information by enabling evaluators to efficiently analyse large document collections. It combines a qualitative appraisal of meaning and complex statistical approaches to extract relevant information from text.

Evidently, text mining is not a fully automated process in which a machine is fed documents and returns meaningful information. Rather, it requires a thoughtfully devised model based on explicated assumptions, a qualitative appraisal of key variables and results, and a critical validation of the algorithms. Deriving meaningful estimates from data presupposes clearly devised concepts and a model of causal assumptions (see box). A combination of the qualitative skills of interpretation and modelling with automated text analysis, makes available the wealth of available information to contribute to evidence-based policy-making.

Process of Text Mining



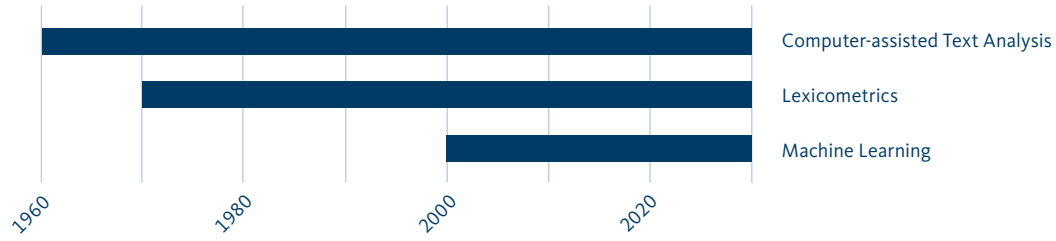
Text mining holds great potential for evaluation. However, some challenges need to be addressed: data acquisition and management can be expensive, long workflows can be time consuming, and initial development requires programming skills. Fortunately, solutions to overcome most obstacles already exist. Especially the availability of open-source software and reusable algorithms that save time and money. Nonetheless, more complex applications and data infrastructures might require cooperation with experienced data scientists.

What is text mining?

Text mining is a set of "computer-based methods for a semantic analysis of text that help to automatically, or semi-automatically, structure text, particularly very large amounts of text" (Heyer, 2009, p.2).

Interest in automated procedures to handle unstructured data goes back at least to the 1950s when the first texts were digitized. Software for computer-assisted content analysis became available in the 1960s and has grown significantly since the 1980s. Software for qualitative text analysis, such as MaxQDA or ATLAS.ti, is now

Figure 1: The Evolution of Text Mining



Source: Authors' own diagram.

widely used in evaluation. First applications of lexicometrics were developed in the 1970s, but until after 2000 there were only few machine-learning approaches to text mining.

Text mining has matured in fields as diverse as biostatistics and computational humanities and is currently spreading to other research areas. The vast availability of digital text as well as software (see box) and algorithms to retrieve and extract meaning from text have led to an increased application of text mining. However, applications in the field of evaluation are still scarce.

Text Mining Software

Open Source

- R (tm package, quanteda)
- Leipzig Corpus Miner
- StanfordCoreNLP
- SpaCy
- Rapid Miner
- GATE
- openNLP

Commercial

- IBM SPSS Text Analytics
- SAS Text Miner
- discovertext
- Microsoft Azure Text Analytics

Text as data

Text mining requires preprocessing. Words, the carriers of meaning, are identified and transformed into a processable data structure. This typically includes the following steps:

- Identification of relevant unstructured text in the sources and separation of sentences and words (tokenization).
- Processing and transformation of the extracted words. Wordforms are mapped to their lexical baseform (lemmatization) and capital letters are transformed to lowercase to unify the vocabulary and to reduce the heterogeneity of semantically similar words.
- Identification of entities (places, names or companies) or part-of-speech information (noun, verb etc.) to enrich the text with information (natural language processing).

This leads to a final data object of separated words, i.e. tokens. A single token is an instance of a unique wordform, a so-called type. The set of all types forms the vocabulary of a document or document collection. Documents are thus stored as sequences of types. These sequences are transformed into matrices and are the key to the statistical processing of textual data.

Methods

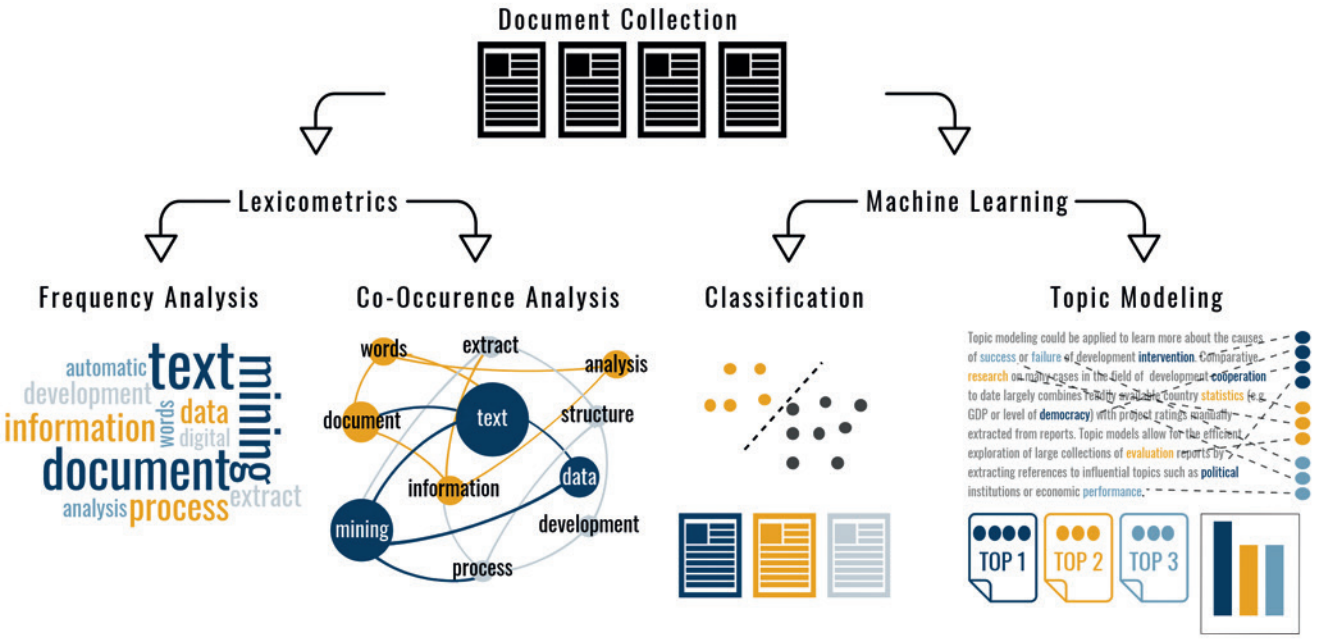
Text mining is a mixed-methods approach as it integrates quantitative and qualitative reasoning. If we observe only the occurrence of symbols (e.g. words), this results in quantitative properties such as frequencies. More advanced approaches also extract context by analysing multiple symbols and their interaction. This allows for a qualitative but still a summarising look into text collections.

We can generally differentiate between lexicometrics and machine-learning approaches.

Lexicometrics, such as frequency and co-occurrence analysis, describes the statistical study of vocabulary. Its main goal is to extract structures from texts for subsequent interpretation.

Frequency analysis: The counting of events is one of the basic operations in text mining. Only the numerical representations of the documents are needed. For example, frequency analysis determines how often a word (or type) occurs within documents. Including meta-data allows for faceting counts by timestamps or authors for time series or comparative analyses.

Figure 2: Text Mining Approaches



Source: Authors' own diagram.

A popular application of frequency analysis is sentiment analysis. It counts positive or negative terms in documents. In development cooperation evaluations, sentiment analysis could be applied to measure public opinion by investigating social media. Additional spatial or temporal information allows for comparing countries or for identifying trends.

Co-occurrence analysis: The co-occurring appearance of words within sentences, paragraphs or documents gives us an idea of their semantical embedding and meaning. It is possible to count the frequency of word co-occurrences and to identify those pairs which appear together more often than expected by chance.

Co-occurrence analysis is well suited to examine how the discourse on a certain topic develops over time. For instance, we might identify words from newspaper reports that typically co-occur with the term "development cooperation". This would allow us to examine whether the discourse changes over time or in reaction to certain events.

Machine learning uses algorithms to learn properties from text examples. These properties can then be predicted in other documents or otherwise explored.

Classification: Classification maps properties of documents (e.g. words) to classes. A trained classifier automatically assigns these classes to unknown text, which allows for the automatic annotation and labelling of text. For example, a text can be classified word-wise in order to identify named entities such as organisations, locations or personal names.

A possible application is geo-referencing of development cooperation projects. Whereas current approaches to geo-coding largely rely on close reading, natural language processing provides an approach to automatically identify locations from text. This provides a clearer picture on distributional patterns of aid within partner countries.

Topic modelling: Topic modelling assigns documents to one or multiple topics. Topics are modelled as distributions over the vocabulary, i.e. topics are clusters of similar words. Documents can thus be summarised, selected and analysed as a mixture of topics.

Topic modelling could be applied to investigate the causes of successful development interventions. Comparative research on many cases to date largely combines available country statistics (e.g. gross domestic product or level of democracy) with project ratings manually extracted from reports. Topic models facilitate

the efficient exploration of large collections of evaluation reports by extracting references to influential topics such as political institutions or economic performance.

Outlook

To date, the number of text mining applications in the study and evaluation of development cooperation remains very limited. Yet given the potential, organisations in this domain have stated their willingness to apply text mining in future evaluations (UN Global Pulse, 2016).

Looking ahead, the number of applications is likely to rise for three reasons. First, the availability of digital text through digitisation of paper sources, digital publishing and social media increases the scope of possible applications. The sheer volume of information makes it necessary to move beyond approaches that exclusively apply 'close reading'.

Second, best practices are evolving due to growing practical experience. Consequently, the field of research has been moving from exploring methodological questions (e.g. which algorithms work best) to applying methods that have demonstrated their validity and reliability. Increasing numbers of tested and reusable algorithms greatly facilitate the application of text mining.

Finally, recent developments promise an increase in practical relevance. The advances in Deep Learning allow for applications in semantic embedding, automatic translation or entity linking and resolution for automatic text understanding and reasoning. Furthermore, the detection of relations and arguments in text becomes possible. All these advances may be studied in the context development report assessment in order to cope with the increasing amount of demanding text reports.

Text mining provides a way to efficiently extract meaningful information from the increasing amounts of text. Its combination of interpretative appraisal and statistical techniques has the potential to generate novel insights, ultimately contributing to evidence-based policy-making. Today, text mining is ripe for application in evaluation.

References

UN Global Pulse (2016), *Integrating Big Data into the Monitoring and Evaluation of Development Programmes.*

Heyer, G. (2009), 'Introduction', In *Text Mining Services – Building and applying text mining based service infrastructures in research and industry. Proceedings of the Conference on Text Mining Services – TMS 2009 at Leipzig University.* Leipzig: LIV.

Dr Andreas Niekler
Computer Scientist
University of Leipzig

Dr Thomas Wencker
Evaluator

The German Institute for Development Evaluation (DEval) is mandated by the German Federal Ministry for Economic Cooperation and Development (BMZ) to independently analyse and assess German development interventions. Evaluation reports contribute to the transparency of development results and provide policy-makers with evidence and lessons learned, based on which they can shape and improve their development policies.

German Institute
for Development Evaluation (DEval)

Fritz-Schäffer-Straße 26 · D-53113 Bonn, Germany
Phone +49 228 33 69 07-0

info@DEval.org
www.DEval.org