Mining big data with computational methods

Annie Waldherr, Gerhard Heyer, Patrick Jähnichen, Andreas Niekler, & Gregor Wiedemann

To cite: Waldherr, A., Heyer, G., Jähnichen, P., Niekler, A., Wiedemann, G. (2016): Mining big data with computational methods. In G. Vowe & P. Henn (Eds.), *Political Communication in the Online World: Theoretical Approaches and Research Designs* (pp. 201-217). New York, NY: Routledge

Abstract

When studying online communication, researchers are confronted with vast amounts of unstructured text data and experience severe limitations to the established methods of manual quantitative content analysis. Text mining methods developed in computational natural language processing (NLP) allow the automatic capture of semantics in massive populations of texts. In this chapter, we present state of the art NLP methods and discuss potential applications and limitations for communication research. Unsupervised methods such as co-occurrence analysis or topic modeling enhance explorative research whereas supervised methods such as machine-learning classification support a deductive research strategy similar to traditional content analysis.

Keywords

Text Mining, Big Data, Web Content, Natural Language Processing, Co-occurrence Analysis, Topic Model, Sentiment Analysis, Machine-Learning Classification, Information Extraction

Introduction

The age of big data poses enormous challenges to traditional methods of empirical research (boyd & Crawford, 2012; Mahrt & Scharkow, 2013; Tinati, Halford, Carr, & Pope, 2014). This is experienced everyday by communication researchers who seek to analyze political discourses online. They have to deal with a seemingly endless amount of web sources. Content is produced at an ever increasing rate leading to massive amounts of text documents to be possibly analyzed. Furthermore, texts are mostly unstructured and available in a variety of formats such as web pages, blog posts or tweets. We experience an increasing volume, velocity, and variety of information which Laney (2001) summarized as the three Vs of big data. Additionally, online content is growing more complex as it is interconnected through hyperlinks or hashtags (Allen Booz Hamilton, 2013).

For this kind of web data, it is increasingly difficult to apply established methods of manual quantitative content analysis. Problems start with defining units of analysis, identifying basic populations, or drawing representative samples before even proceeding to coding the data (for more challenges to online content analysis see chapter 11).

A different approach to analyzing text has been developed in computational *natural language processing* (NLP) (Feldman & Sanger, 2006; Heyer, Quasthoff, & Wittig, 2006; Manning & Schütze, 2003). NLP text mining methods allow to automatically capture semantics of texts in unstructured corpora. Massive populations of text documents can be analyzed with limited effort so that drawing restrictive samples is no longer necessary. Not surprisingly, more and more communication scholars explore the possibilities of these computational methods (e.g., Scharkow, 2013; Stieglitz, Dang-Xuan, Bruns, & Neuberger, 2014; van Atteveldt, Kleinnijenhuis, & Ruigrok, 2008).

The aim of this chapter is to present the state of the art in text mining methods and discuss potential applications and limitations for political communication research. We proceed by introducing the basic approach of NLP in section 2, before we present an overview on specific methods of text mining in section 3. We conclude by discussing implications and future perspectives for political communication research.

Text mining and models of semantics

The broad set of methods to semantically structure (very) large amounts of unstructured text data is referred to as *text mining*.[1] A crucial decision for text mining applications is how to model semantics of text. Turney and Pantel (2010, p. 141) refer to semantics as “the meaning of a word, a phrase, a sentence, or any text in human language, and the study of such meaning”. In NLP, three types of semantic processing models may be distinguished: patterns of character strings, logical representations of entity relations, and distributional semantics.

Patterns of character strings

In computational environments, text is basically represented by character strings as primary data format. The simplest model to process meaning is to look for predefined patterns in these character sequences. Imagine for example the sequence “United States” occurring in a text document as representing ‘the country United States of America’. By extending this single sequence to a set of sequences, e.g. {“United States”, “Germany”, “Ghana”, “Israel”, …}, we create a reference to ‘a country’. Such lists of character sequences representing meaningful concepts, also called *dictionaries*, have a long tradition in communication science (Stone, Dunphy, Smith, & Ogilvie, 1966). By using a formal language for search and replace operations (regular expressions) and elaborated dictionaries it is possible to model very complex concepts even with this rather simplistic approach. In practice, however, success of this approach heavily depends on the skills and experience of the researcher developing such dictionaries.

Logical representations of entity relations

A much more ambitious approach to processing semantics is the employment of logic frameworks, e.g., first-order logic or description logics such as OWL[2], to model relations between semantic units represented by linguistic patterns. Logical connectives and quantifiers are used to combine such units into a knowledge base, also called formal ontology, which allows for reasoning. As basic example imagine a set of two rules 1) ***x*** *is a red car*, and 2) *all*

[1] This is the main difference to data mining. While data mining methods can only be applied to extract knowledge from structured data in databases text mining methods apply to unstructured text.

[2] OWL - Web Ontology Language, see http://www.w3.org/TR/owl2-overview/

cars are vehicles as a formal ontology. Then, querying for *all red vehicles* would yield the result ***x***, although the knowledge base only contains explicit information about the red car ***x***. Setting up a formal set of rules and connections of units in a complete and coherent way, however, is a time consuming and complex endeavor. Probabilistic models for automatic semantic parsing can be utilized to support generation of such rule sets (Beltagy, Erk, & Mooney, 2014). But, up to now quality and level of granularity of such knowledge bases are insufficient for many practical applications.

Distributional semantics

Distributional approaches to processing semantics are based on the *'bag of words'* assumption that frequencies of terms in a document mainly indicate the meaning of its content; i.e., "words that occur in similar contexts tend to have similar meanings" (Turney & Pantel, 2010, p. 148). Order of terms in contrast is less important and can be disregarded. This is certainly not true for most human real world communication, but works surprisingly well for many NLP applications.[3]

The *vector space model* (VSM), utilized initially for Information Retrieval (Salton, Wong, & Yang, 1975), encodes counts of occurrences of single terms in documents (or other context units such as sentences) in vectors of the length of the entire vocabulary of a modeled corpus. If there are *M* different word types in a collection of *N* documents, then the counts of *M* word types in each of the documents leads to *N* vectors which can be combined into a *N x M* matrix, a so-called document-term-matrix (DTM).

The construction of a DTM usually is achieved with a sequential process chain called *preprocessing*. First, sentences and single terms (tokens) are identified before eventually deleting certain tokens, so-called stop words,[4] which do not contribute much to the meaning of a text. Furthermore, it may be useful to unify variants of terms expressing the same meaning by stemming (removal of suffixes of terms by language specific rules) or lemmatization (transformation of inflected forms to their dictionary form).

[3] The complete loss of information on word order can be mitigated by observing n-grams, i.e., concatenated ongoing sequences of n terms instead of single terms while creating a DTM.

[4] Stop words are high-frequent, functional terms in language unspecific to a certain topic, e.g., *a, the, is, are, have*. For many NLP applications it is recommendable to ignore them.

For online sources initial extraction and cleaning steps are necessary, often referred to as *web scraping* (Munzert, Rubba, Meißner, & Nyhuis, 2014). This encompasses the task to identify the relevant textual parts from a crawled HTML. Content units such as title and body text of a blog article have to be separated from menu link texts, sidebar content or advertisements. Depending on the structure of the web page this can be a very tricky task involving lots of heuristics.

Once a document collection is encoded in a numerical DTM format, it can be used as an input to many NLP applications. In the following, we introduce some applications that promise to be useful particularly for online political communication research. Thereby we mainly focus on methods from the approach of distributional semantics because these have proven to be most efficient.

Text-mining applications

As known from data mining, we distinguish unsupervised from supervised methods for data analysis. While unsupervised methods help to explore structures in large amounts of unknown data, supervised methods take into account external knowledge to train machine-learning algorithms.

Unsupervised methods

The following methods are data-driven approaches identifying previously unknown patterns and structures emerging from the data itself. Thus, they support an inductive research strategy.

Term extraction

For any document, or collection of documents, we can identify key terms by applying statistical measures (Archer, 2009). The method of so-called *difference analysis* compares the frequencies of terms in the target corpus to frequencies in a reference corpus consisting of general texts of the same language without a bias to any topic.[5] Comparisons to more specific reference corpora are also possible. For example, if we are interested in comparing agendas

[5] Large collections of textual data such as the data of the Leipzig Corpora Collection (LCC) have proven to be quite suitable for that (Biemann, Heyer, Quasthoff, & Richter, 2007).

and positions of two politicians or parties, we can directly compare corpora consisting of their speeches with respect to the key terms used. Scharloth (2013) conducted such an analysis to reveal differences of language use between candidates Angela Merkel and Peer Steinbrück during the campaign for German federal elections.

Difference in term usage is based on a statistical test that measures the amount of surprise when observing term frequencies in the target text with respect to the reference text after normalizing the overall size of both texts. Dunning's log-likelihood ratio test (Dunning, 1993) has proven to deliver good results. In effect, we get a ranked list of terms that significantly differ in their frequency with respect to the target and the reference text (Rayson & Garside, 2000).[6]

Analysis of significant co-occurrences

Although speakers usually assume that single terms already express a meaning on their own, following the approach of distributional semantics, this meaning should be seen as a function of surrounding contexts in the overall system of language. This can be achieved by evaluating *significant co-occurrences* of words within texts on the level of whole documents, paragraphs, or sentences.

Instead of just counting frequencies of co-occurring terms, co-occurrence analysis calculates the statistical significance of each and every co-occurring pair of words. This approach is based on the assumption that only pairs of words that exhibit a significant joint occurrence within a corpus indicate a salient context of usage. In practice, several statistical measures for co-occurrence significance can be used (Heyer et al., 2006), but for most cases Dunning's log-likelihood has again proven to deliver best results (Bordag, 2008).

Having selected a word of interest, its significant co-occurrences can be depicted as a network of word usages or a list of all significant co-occurrences (see figures 1 and 2 for examples). Visualizing its meaningful interactions with other words in a given collection of texts results in a global view on the semantic context of a word (Heyer et al., 2006, p. 134 ff.). Co-occurrences of different sub-collections of a corpus can also be analyzed comparatively to reveal semantic changes across time, space, sources, or topics.

[6] The software WordSmith tools (http://www.lexically.net/wordsmith) provides a well known implementation of this term extraction method.

From the perspective of communication research such co-occurrence networks may be interpreted as frames in the sense of likely associations or interpretations (e.g., Miller, 1997; Hellsten, Dawson, & Leydesdorff, 2010). Van Atteveldt (2008) has developed a similar approach to measure so-called "associative frames". He calculates the conditional probability of one concept occurring in the context of another concept and depicts these relationships in semantic networks. However, focusing only on conditional probabilities and not on significance emphasizes terms that are highly dependent from each other, but not necessarily terms that occur more frequently in the corpus.

Topic Models

Another unsupervised method that makes use of co-occurring words in documents is topic modeling (Blei, Ng, & Jordan, 2003; Steyvers & Griffiths, 2005). A topic model is a Bayesian hierarchical probabilistic (graphical) model. It defines an artificial document model describing how the words in the documents get into their place. Instead of using a frequentist approach (as in co-occurrence analysis above) we adopt a Bayesian approach: We suggest an initial guess about the structure of the model, the *prior*, and then define the likelihood of the data under a certain model structure.

In a topic model, two *latent factors* form this structure and may be interpreted as (1) the *topics* themselves, and (2) the *documents' topic proportions*. Our prior belief (the initial guess) about document collection structures in general is the following: We assume that each topic (to be understood as a semantic class of words) will be characterized by only a small subset of the vocabulary. In turn, we also expect only a few of the hypothetical topics to be present in each document. The appropriate prior distribution for both factors is the Dirichlet distribution.[7]

By updating the prior under the influence of data, keeping the data likelihood high, we arrive at a *posterior* belief about the model structure. We can now explore the posterior distribution and derive sets of words as semantic categories. Note that the connections between words identified by topic models are *latent*, i.e., not observed directly. This is a distinctive feature compared to co-occurrence analysis. Using a topic model, we can reveal a latent semantic

[7] A prior distribution is a distribution that produces the distribution of interest from a random draw. In our case, a draw from the Dirichlet distribution produces a multinomial, i.e. a point on the simplex.

connection between words, even if they never occurred in a document together. The connection is simply built by other terms both words have co-occurred with across the document set.

We can also use this posterior belief to make inferential statements about previously unseen data. This is the key benefit of this line of thinking. Using word co-occurrence analysis alone, we could never deduce any information about a pair of terms that was previously unseen in the referential corpus. Note, however, that the probabilistic nature of topic models demands a thorough inspection of model outcome and checking of the models in use (cf. Gelman & Shalizi, 2013).

To exemplify, we examined the State of the Union addresses and speeches to congress held by the US-President since 1900 which are publicly available online (Woolley & Peters, 1999). We present selected findings related to the key terms “soviet” and “terror”. Figures 1 and 2 show co-occurrence networks for both those terms respectively. The thicker an edge, the more significant a co-occurrence is (i.e., we observe such a pairing considerably more often than we would expect by chance). Strongly significant edges are additionally color coded in orange. Figure 3 shows a comparison of the relative word frequencies for both terms, i.e., their relative frequency proportions over time. Finally, figure 4 shows an example outcome of a topic model on the data described. We selected topics that contain either “terror” or “soviet” in their top 25 words (when sorted by probability in that topic).

-- Insert figures 1-4 approximately here --

Quite interestingly, considering co-occurrence and relative frequency analysis alone, we would deduce that the President stopped talking about “soviet” and right after that started to talk about “terror” in the same manner as he spoke about “soviet” before. Both terms show similar word co-occurrences with terms such as “military”, “forces”, “security”, “defense” or “free”. Mentioning of the term “soviet” ceased in the beginning of the 1990s when the Soviet Union dissolved. The term “terror” experienced a surge at the beginning of the 2000s just after the 9/11 attacks.

However, adding topic model analysis, we see that “soviet” and “terror” were *not* discussed in the same way (see figure 4). Topic 8, 22 and 35 show the usage of “soviet” in different semantic contexts. Topic 8 is about the Non-Proliferation Treaty between the Soviet Union and the US; Topic 22 and 35 are about US concerns about communism in the world in general that is assumed to pose a threat. Here the term “soviet” is not central to the debate but appears as just one of the communist nations talked about. This helps us to distinguish the different peaks in relative frequency of “soviet” that we observe in figure 3. On the contrary, Topic 10 about “terror” is clearly confined to the concept of terror as coined since 9/11, also as suggested by figure 3. We can deduce that neglecting the semantic category a certain word co-occurrence belongs to, may lead us to false conclusions. With topic modeling we have a ready-to-use tool to enrich our findings semantically.

Communication researchers are just about to discover the opportunities of topic modeling. First of all, topic models are a promising way of estimating the salience of topics in the corpus – a classic task of manual content analysis. For example, Evans (2013) uses topic modeling to identify issues discussed as “unscientific” in American newspapers between 1980 and 2012. However, the clusters of semantically connected words a topic model identifies need intense inspection and interpretation on the part of the researcher. Whether it is appropriate to interpret them as substantive topics or issues in the sense of political communication theory cannot be guaranteed. In some cases, the word clusters might describe more specific sub-topics, or even frames (Maier, Waldherr, Miltner, Jähnichen, & Pfetsch, 2015; van Atteveldt, Welbers, Jacobi, & Vliegenthart, 2014).

Besides the analysis of topical structures of text corpora themselves, topic models might also be helpful in earlier stages of the research processes, e.g., identifying populations of relevant text documents on an issue. A keyword-based search in a document collection may be enhanced with a topic-based exploratory search that can recommend documents with similar thematic structure. This is particularly helpful if the issue of interest cannot be fully described by a catchy keyword or phrase.

Supervised Methods

Supervised methods in machine learning (ML) rely on the inclusion of external knowledge to infer models on the data. This external knowledge usually is a set of categories and assignments of these categories to a set of training data entities, e.g., documents. Based on this knowledge we can decide to which category a new, so far unobserved document belongs to. This process is called *text classification* and is a deeply investigated problem in NLP. It can be useful for a variety of purposes. A well-known application to almost every internet user are spam detection systems which automatically identify junk emails.

For text classification different ML algorithms have been successfully used. Two of the most common approaches are *Naive Bayes* (NB) and *Support Vector Machines* (SVM; Joachims, 1998). For each document, both algorithms provide a decision of either 0 or 1 whether a document belongs to a category or not. For coding systems of more than one category, the process can be modified to enable *multiclass classification* (exactly one label needs to be selected for each document), or *multi-label classification* (one or more labels could be selected for each document).

Document classification for content analysis

An interesting ML application for communication research is the classification of whole documents such as newspaper articles into thematic categories. Scharkow has conducted exemplary studies with both SVM (2012) and NB (2013) algorithms. He showed that ML works to classify newspaper articles into rather rough categories such as economy, sports, interior and foreign politics.

The supervised process of ML text classification resembles the manual process of content analysis. The ML classifier first is trained on a manually coded sample to learn to correctly assign predefined codes to documents. The process infers the coding rules on its own by identifying discriminating features for each category from the training data. To make this work, training data needs to be coherent, complete and disjoint with respect to the classes defined, i.e., at least one and not more than one category definition must apply to every context unit. Also, training data should include as much variety as possible for any category.

Besides a considerate process of training data generation, several adjustments can be made to optimize ML classification. These include feature engineering[8] and optimal feature selection strategies[9]. The resulting ML algorithm may be considered as a trained ‘naive coder’ which can now be applied to any (sub-)set of text collections comparable to the training data set.

[8] Feature engineering includes decisions about the information needed to accurately identify categories: Are word counts sufficient, or do I need more information such as combinations of word types or syntactical features?

[9] In optimal feature selection we decide, which of the extracted word or syntax features are discriminative for a category.

After classification, for each document of the collection we have a decision whether it belongs to a category or not. This allows for an evaluation similar to inter-coder reliability tests. Assuming a large set of training data, we can split this into two halves and train the ML algorithm on the first set. The second set is then used for automatic classification. Now we can compare the predicted and the actual labels of the documents, and assess *precision* (share of correctly identified positive labels for a certain category) and *recall* (share of positively identified documents on all existing documents of a certain category). We can also compute reliability measures such as Cohen's Kappa between human and ML raters. As training data most often is rare, manually coded sample sets are usually not split into halves but into *k* folds for *k*-fold cross validation.

Active learning procedures can help to compile optimal training data sets that require less examples, but provide higher classification accuracy (Settles, 2010). Lemke, Niekler, Schaal, and Wiedemann (2015) applied such a process to classify paragraphs from newspaper articles containing the category "economized justification of politics". Training of a SVM classifier was initialized by a manual set of 220 paragraphs that had been identified as good examples for the category of interest. This set was then augmented in seven iterated active learning runs of the classifier, each run providing 200 new paragraphs with a positive classification of new unknown texts. Manual evaluation of these 200 results by the research team lead to new high quality positive and negative examples to enrich the training set. The final training set after seven iterations consisted of 653 positive and 1,749 negative sample paragraphs, resulting in a F1-measure = .613 for 10-fold cross validation on this training set.[10] Similar to Cohen's Kappa, or Krippendorff's alpha, one strives for values of .7 or above. For complex content categories this might be hard to achieve, but it is actually accomplished for more clear-cut distinctions.

For example, Colleoni, Rozza, and Arvidsson (2014) successfully trained an ML classifier to (1) identify political tweets in a US Twitter corpus and (2) distinguish Democrat vs. Republican political orientation of the tweets. They report an F1-measure of .79 and higher for 10-fold cross-validation on their training set. Their study also shows that it is possible to apply text classification to smaller context units such as short sentences. However,

[10] The F1-measure is the harmonic mean between precision and recall defined above. It ranges between 0 and 1.

it has to be considered that these units provide rather little information to an ML algorithm which makes it generally more complicated to train an efficient algorithm.

Sentiment analysis

Another example application for supervised classification is *sentiment analysis*, the identification of subjective information or attitudes in texts (Pang & Lee, 2008). It may be realized as a ML classification task assigning either a positive, neutral, or negative class to a document set narrowed down to a specific context beforehand (e.g., using a topic model). Classification then allows for the tracking of attitudes in these documents over time in a reliable way.

Especially text data of online communication became of recent interest for automatic detection of sentiments in election contexts. Johnson, Shukla, and Shukla (2011) analyzed around 550,000 twitter posts on Barack Obama and cross-correlated their findings with national survey data on popularity of the president. Their findings suggest that short term events affecting Twitter sentiments do not necessarily affect president's popularity significantly. Tumasjan, Sprenger, Sandner, and Welpe (2010) computed plausible sentiment profiles of politicians and parties of the German parliamentary elections in 2010 by analyzing more than 100,000 tweets. Interestingly, they also found that mere frequency of mentioning major parties pretty accurately predicted election results.

Information extraction

As mentioned above, supervised classification not only works for complete documents. It also applies to single terms or sequences of terms fitting into a certain category. Sequence classifications such as part-of-speech tagging, syntactic parsing or named entity recognition (NER) leave behind the “bag of words” assumption by taking local context terms into account. These procedures are not useful for political communication analysis as such. They rather constitute useful preprocessing steps to filter desired contexts for analysis. Part-of-speech tagging for instance can be used to filter document contents for certain word types before any subsequent text mining application. Term extraction or topic models then can be realized by just concentrating on nouns or verbs.

Syntactic parsing splits sentences into coherent sub-parts and reveals their syntactic relations. This may be applied to identify desired subject-object relations (“In America, you watch Big Brother.” versus “In Soviet Russia, Big Brother watches you!”) or to build discriminating features for document classification. Kleinnijenhuis and van Atteveldt (2014) use parsing information on news coverage of the middle east conflict to distinguish speech acts expressing Israel as an aggressor against Palestine or vice versa.

Last but not least, named entities (such as person names, organizations, or locations) can be extracted and classified to identify actors in texts.[11] These can then be related to structures of extracted meaning such as certain language use measured by significant term extraction.

Conclusion

This short overview has shown that communication scholars can immensely profit by opening up to computational methods of text mining based on NLP. Computer scientists dispose of an array of suitable tools for the purposes of content analysis. Because they allow to semantically analyze vast corpora of unstructured text, these methods are particularly interesting for political communication researchers studying online content. Apart from the mentioned additional efforts in the preprocessing steps, the methods can be readily applied to online corpora as to any other digital text corpus. However, until now there exist hardly any standard software solutions that are applicable for the ordinary communication researcher without any further technical know-how. Therefore, to date it seems inevitable to strengthen interdisciplinary cooperations with computer scientists.

Table 1 gives an overview on how text mining specifically enhances the traditional toolbox of content analysis. Dictionary approaches (which have not been further elaborated here) and supervised classification are closest to the traditional, deductive logic of quantitative content analysis. However, there (still) are severe limits to the interpretative knowledge and abilities of supervised machine-learning. Up to now, rather complex concepts such as frames have not been coded with sufficient accuracy, although it has to be admitted that these constructs also pose high challenges to inter-coder reliability of human coders. In contrast, the coding of rather broad topics and attitudes (such as sentiments or political orientation) can be

[11] A well-matured reference implementation of a Conditional Random Field approach to Named Entity Recognition is provided by the NLP group of Stanford University (Finkel, Grenager, & Manning, 2005).

successfully delegated to computational algorithms. Also named entity recognition works reliably to identify specific actors or organizations in a text corpus.

Table 1: How text mining contributes to the toolbox of content analysis

	Methodological approach	
Research strategy	**Quantitative**	**Qualitative**
Theory-driven/ deductive	*Quantitative content analysis* Dictionaries Supervised classification	
Data-driven/ inductive	Term extraction Co-occurrence analysis Topic modeling	*Qualitative content analysis*

Note: Traditional methods of manual content analysis are written in italics.

Inductive, unsupervised methods such as significant term extraction, co-occurrence analysis and topic modeling add a completely new approach to the common toolbox of content analysis. While following a quantitative, statistical approach, they are inherently data-driven and inductive. Therefore, they are particularly valuable for exploratory purposes that have been traditionally pursued with manual qualitative content analysis for only small samples. For instance, topic modeling does not search for pre-defined topics, but structures the whole corpus in terms of emerging topic clusters. The same is true for co-occurrence analysis: Unexpected associations of words might appear during analysis. However, in any case the found structures need intensive interpretation as well as plausibility checks. Researchers have to be very familiar with their text corpus including its thematic and temporal context to be able to validly interpret statistical topics as issues or co-occurrences as frames. Otherwise they risk to overinterpret methodological artefacts.

In our view, one of the biggest potentials of text mining approaches lies in the many possibilities of combining different supervised and unsupervised methods (and our overview is far from exhaustive). For example, first identifying actors with named entity recognition and then connecting them to their significant co-occurrences, related topics from a topic model and sentiments from a machine-learning classifier will bring us closer to the end of automated discourse and frame analysis. At least we can answer questions such as: Who says what with which sentiment in what context? And we can answer them not only on the document level, but also on the level of paragraphs and sentences, which draws us near the analysis of claims, statements or arguments.

Of course, computational content analysis cannot be of the same depth as manual analysis. The big advantage of text mining is that we gain an overview on the content of vast text corpora with limited efforts. This is particularly interesting for comparative analyses when we want to juxtapose slices of the text corpus. Here, traditional content analysis entailing sampling and manual coding soon becomes very extensive because samples of sufficient size have to be drawn for every relevant sub-population. When working with large network data researchers might even be interested in the content data of each node (actor) in the network. This is relevant for instance for studying the topology of hyperlink networks on the Internet (see chapter 15) in terms of content: What do people post on the Web and how are they connected? Getting this information for every node in the network would be impossible without relying on automated methods (Maier et al., 2015).

Finally, there are also many possible combinations of automated and manual methods of content analysis. Semi-automated content analysis systems combine the "best of both worlds" (Wettstein, 2014). They interact with human coders, propose plausible codes and continuously learn from their final coding decisions (see also Wueest, Clematide, Bünzli, Laupper, & Frey, 2011). Sometimes text mining might also be helpful for identifying relevant text documents from large data bases to prepare an in-depth manual content analysis (Waldherr, Maier, Miltner, & Günther, 2014).

References

Archer, D. (Ed.). (2009). *What's in a word-list? Investigating word frequency and keyword extraction*. Surrey, UK: Ashgate.

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern information retrieval: The concepts and technology behind search*. Harlow, UK: Addison Wesley.

Beltagy, I., Erk, K., & Mooney, R. (2014). Semantic parsing using distributional semantics and probabilistic logic. *Proceedings of the ACL 2014 Workshop on Semantic Parsing* (pp. 7-11). Stroudsburg, PA: Association for Computational Linguistics.

Biemann, C., Heyer, G., Quasthoff, U., & Richter, M. (2007). The Leipzig Corpora Collection: Monolingual corpora of standard size. *Proceedings of Corpus Linguistic 2007*. Birmingham, UK.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993-1022.

Booz Allen Hamilton (2013). The field guide to data science. Retrieved from http://www.boozallen.com/insights/2013/11/data-science-field-guide

Bordag, S. (2008). A comparison of co-occurrence and similarity measures as simulations of context. *Computational Linguistics and Intelligent Text Processing*, *4919*, 52-63. doi: 10.1007/978-3-540-78135-6_5

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662-679. doi: 10.1080/1369118X.2012.678878

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication, 64*(2), 317-332. doi: 10.1111/jcom.12084

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61-74.

Evans, M. S. (2013). A computational approach to qualitative analysis in large textual datasets. *PLoS ONE*, *9*(2), e87908–e87908. doi:10.1371/journal.pone.0087908

Feldman, R., & Sanger, J. (2006). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge, UK: Cambridge University Press.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43rd Annual*

Meeting of the Association for Computational Linguistics (pp. 363-370). Stroudsburg, PA: Association for Computational Linguistics.

Gelman, A., & Shalizi, C. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66*(1), 8-38. doi: 10.1111/j.2044-8317.2011.02037.x

Harris, Z. (1954). Distributional structure. *Word, 10*(23), 146-162.

Hellsten, I., Dawson, J., & Leydesdorff, L. (2010). Implicit media frames: Automated analysis of public debate on artificial sweeteners. *Public Understanding of Science, 19*(5), 590-608. doi: 10.1177/0963662509343136

Heyer, G., Quasthoff, U., & Wittig, T. (2006). *Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse*: W3L.

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. Retrieved from http://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf

Johnson, C., Shukla, P., & Shukla, S. (2011). On classifying the political sentiment of tweets. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.229.3927

Kleinnijenhuis, J., & van Atteveldt, W. (2014). Political positions and political cleavages in texts. In B. Kaal, I. Maks & A. Van Elfrinkhof (Eds.), *From text to political positions* (pp. 1-20). Philadelphia, PA: John Benjamins.

Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lemke, M., Niekler, A., Schaal, G. S., & Wiedemann, G. (2015). Content analysis between quality and quantity: Fulfilling blended-reading requirements for the social sciences with a scalable text mining infrastructure. *Datenbank-Spektrum, 15*(1), 7-15. doi: 10.1007/s13222-014-0174-x

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media, 57*(1), 20-33. doi: 10.1080/08838151.2012.761700

Maier, D., Waldherr, A., Miltner, P., Jähnichen, P., & Pfetsch, B. (2015). *Exploring issues in a networked public sphere: Combining hyperlink network analysis and topic modeling*. Paper presented at the Annual Conference of the International Communication Association (ICA), San Juan, Puerto Rico.

Manning, C., & Schütze, H. (2003). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Miller, M. M. (1997). Frame mapping and analysis of news coverage of contentious issues. *Social Science Computer Review, 15*(4), 367-378. doi: 10.1177/089443939701500403

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). *Automated data collection with R: A practical guide to web scraping and text mining*. Chichester, UK: Wiley.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundation and trends in information retrieval, 2*(1-2), 1-135. doi: 10.1561/1500000011

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora* (pp. 1-6). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1604683.1604686.

Salton, G., Wong, A., & Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Scharkow, M. (2012). *Automatische Inhaltsanalyse und maschinelles Lernen*. Retrieved from http://underused.org/dissertation/

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity, 47*(2), 761-773. doi: 10.1007/s11135-011-9545-7

Scharloth, J. (2013). Die Rhetorik von Angela Merkel und Peer Steinbrück im Vergleich. Retrieved from http://polittrend.de/politik/blog/?p=210

Settles, B. (2010). Active learning literature survey. Retrieved from http://burrsettles.com/pub/settles.activelearning.pdf

Steyvers, M., & Griffiths, T. L. (2005). Probabilistic topic models. In T. Landauer, D. McNamara & W. Kintsch (Eds.), *Latent semantic analysis: A road to meaning* (pp. 427-448). Mawah, NJ: Lawrence Erlbaum.

Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social media analytics. *Business & Information Systems Engineering, 6*(2), 89-96. doi: 10.1007/s12599-014-0315-7

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A computer approach to content analysis*: MIT Press.

Tinati, R., Halford, S., Carr, L., & Pope, C. (2014). Big data: Methodological challenges and approaches for sociological analysis. *Sociology, 48*(4), 663-681. doi: 10.1177/0038038513511561

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (pp. 178-185).

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141-188. doi:10.1613/jair.2934

van Atteveldt, W. (2008). *Semantic network analysis: Techniques for extracting, representing, and querying media content*. Retrieved from http://dspace.ubvu.vu.nl/handle/1871/15964

van Atteveldt, W., Kleinnijenhuis, J., & Ruigrok, N. (2008). Parsing, semantic networks, and political authority using syntactic analysis to extract semantic relations from Dutch newspaper articles. *Political Analysis, 16*(4), 428-446. doi: 10.1093/pan/mpn006

van Atteveldt, W., Welbers, K., Jacobi, C., & Vliegenthart, R. (2014). LDA models topics… But what are 'topics'? Retrieved from http://vanatteveldt.com/wp-content/uploads/2014_vanatteveldt_glasgowbigdata_topics.pdf,

Waldherr, A., Maier, D., Miltner, P., & Günther, E. (2014). *Big data, big noise: The challenge of extracting issue networks from the web*. Paper presented at the Annual Conference of the International Communication Association (ICA), Seattle, WA.

Wettstein, M. (2014). 'Best of both worlds': Die halbautomatische Inhaltsanalyse. In K. Sommer, M. Wettstein, W. Wirth & J. Matthes (Eds.), *Automatisierung in der Inhaltsanalyse* (pp. 16-39). Köln: von Halem.

Woolley, J., & Peters, G. (1999). *The American Presidency Project*. Retrieved from http://www.presidency.ucsb.edu/.

Wueest, B., Clematide, S., Bünzli, A., Laupper, D., & Frey, T. (2011). Electoral campaigns and relation mining: Extracting semantic network data from newspaper articles. *Journal of Information Technology & Politics, 8*(4), 444-463. doi: 10.1080/19331681.2011.567387

Figures

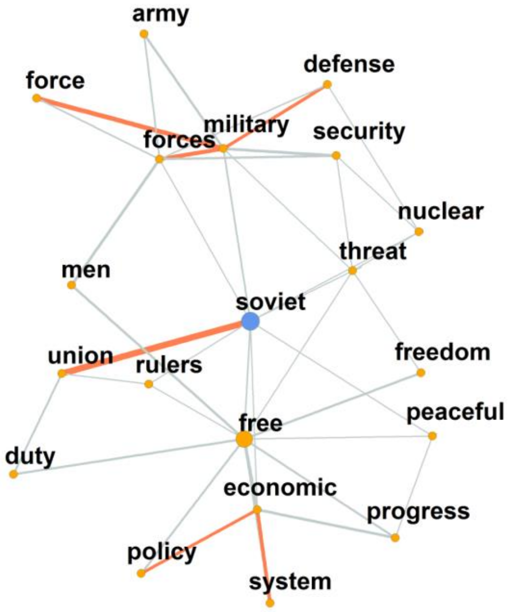


Figure 1. Co-occurence graph for the word *soviet*. The graph is based on the State-of-the-Union-Address corpus.

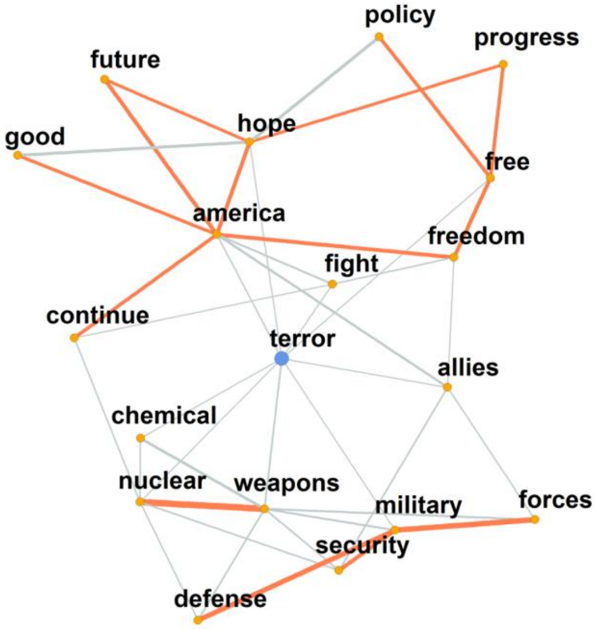


*Figure 2***.** Co-occurence graph for the word *terror*. The graph is based on the State-of-the-Union-Address corpus.

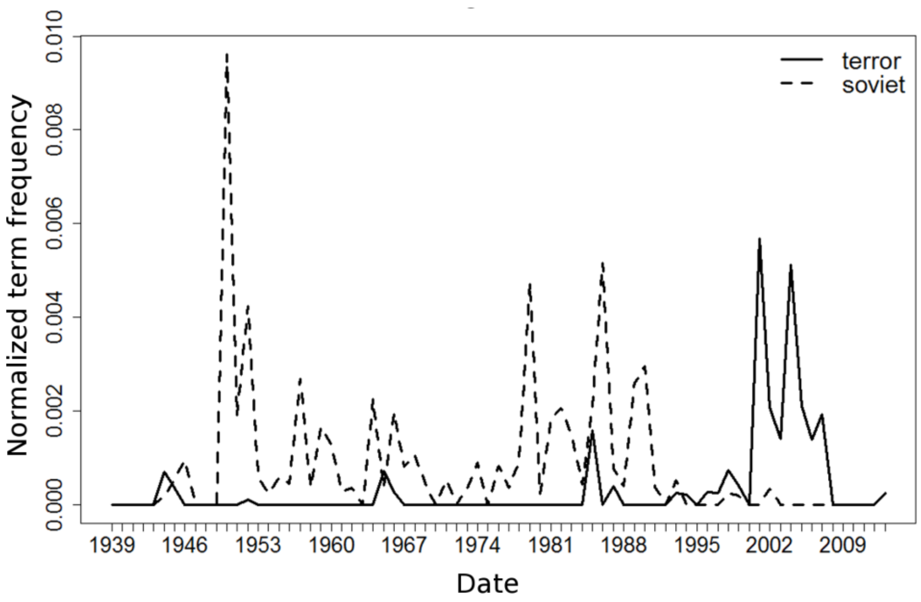
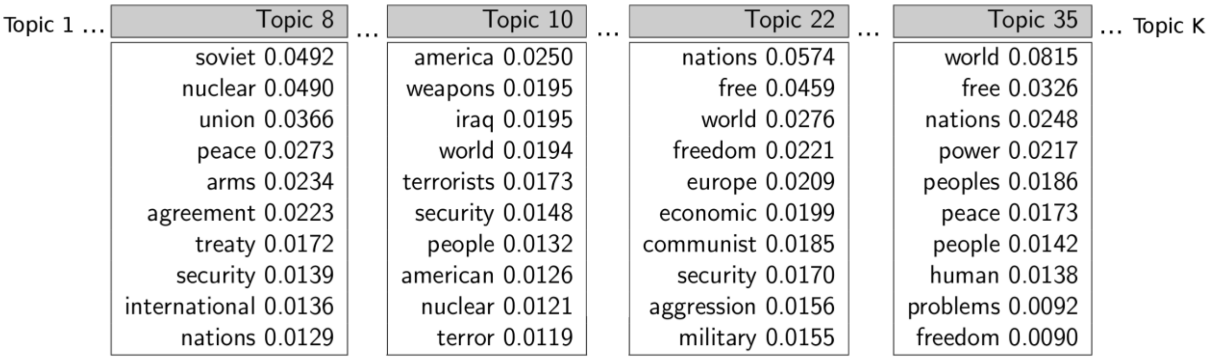


Figure 3. Frequency plot for the words *soviet* and *terror*. The frequencies are normalized for each year by the yearly total amount of tokens.



Topic 1 …

Topic 8	…	Topic 10	…	Topic 22	…	Topic 35
soviet 0.0492		america 0.0250		nations 0.0574		world 0.0815
nuclear 0.0490		weapons 0.0195		free 0.0459		free 0.0326
union 0.0366		iraq 0.0195		world 0.0276		nations 0.0248
peace 0.0273		world 0.0194		freedom 0.0221		power 0.0217
arms 0.0234		terrorists 0.0173		europe 0.0209		peoples 0.0186
agreement 0.0223		security 0.0148		economic 0.0199		peace 0.0173
treaty 0.0172		people 0.0132		communist 0.0185		people 0.0142
security 0.0139		american 0.0126		security 0.0170		human 0.0138
international 0.0136		nuclear 0.0121		aggression 0.0156		problems 0.0092
nations 0.0129		terror 0.0119		military 0.0155		freedom 0.0090

… Topic K

Figure 4. Sample topics created by an LDA model of 50 topics. The model is based on the State-of-the-union-address corpus confined to documents since 1900. Model parameters have been fitted to the data.