This is a preprint of a chapter written for the *Handbook of Quantitative Research Methods in Communication Science*, edited by Lijiang Shen (Pennsylvania State University) and to be published by De Gruyter Mouton.

Computational text analysis

Marko Bachl (Freie Universität Berlin; marko.bachl@fu-berlin.de) and *Michael Scharkow* (Johannes Gutenberg-Universität Mainz; scharkow@uni-mainz.de)

1. Introduction[1]

Computational text analysis (CTA) comprises techniques for measuring the content of texts with the help of computer algorithms. The methods are discussed under various labels, such as text-as-data, automated content analysis, natural language processing, or text mining. The defining characteristic of a CTA technique is that once it is initially configured, the computational system performs the measurements independently without requiring any manual intervention or effort. The strength of CTA lies in its scalability, enabling the measurement of characteristics across vast amounts of text. As a result, CTA has seen widespread application in communication, related social sciences, and the digital humanities, with the increasing availability of digital or digitized, machine-readable texts.

We start this chapter with an overview of the historical development of CTA. We then systematize CTA along two dimensions: the representations of texts for the computational analysis and the supervision of the measurement process. While doing so, we provide some examples of popular techniques. The chapter ends with an outlook into the near future.

2. Historical development

The history and development of CTA can be understood through three central themes: (1) the conceptual complexity of content analytical measures and the potential for measuring them with computer algorithms, (2) the development of hardware and software resources for CTA, and (3) the availability of digital, machine-readable texts. The first phase of CTA development in the late 1950s was characterized mainly by experiments using the computer as a new tool for the social sciences. These initial experiments were almost exclusively limited to producing text statistics, such as word counts, a technique that had been applied in disciplines including political science and literature studies since the 1920s (Holsti, 1969; Stone et al., 1966). Conceptually, these simple CTA methods were much less complex than contemporary manual content analysis approaches (Lasswell et al., 1952; Osgood, 1959).

However, early content analysts were increasingly troubled by the costs of manual classification. They had high expectations for CTA, viewing its development as central to the success of content analysis as a method for the social sciences (Stone, 1997). However, CTA was plagued with numerous problems. Hardware and software limitations allowed only small amounts of text to be analyzed and supported only a few simple measures (Iker & Harway,

[1] This chapter is a completely revised and extended version of Scharkow (2017).

1965). Additionally, all texts had to be digitized through a complex and error-prone process using punch cards. Thus, in the early 1960s, CTA was neither cost-effective—with half an hour of computing time costing as much as a secretary's monthly salary (Stone, 1997, p. 42)—nor less laborious than traditional content analysis.

The introduction of the *General Inquirer* (Stone et al., 1966) and *Words* (Iker & Harway, 1965) programs in the 1960s marked a milestone in the development of CTA. These software packages enabled CTA with relative ease, representing the prototypes of two techniques—dictionary-based classification and co-occurrence analysis—that would dominate the discipline for several decades. Despite continued technological advancements in the 1970s that made computers more accessible to social scientists, conceptual and methodological development of CTA slowed, and interest in these techniques waned, especially in the United States. However, many German scholars continued the research program of the *General Inquirer* by developing dictionaries and software such as *Textpack*, which were primarily designed for classifying open-ended survey questions but also used for text analysis (Züll & Mohler, 2001).

The lack of machine-readable documents remained a limitation for many studies until the late 1970s, restricting researchers to archive material or non-news media. In 1973, DeWeese pioneered the automatic collection of daily media content by leveraging the increasingly common electronic typesetting machines used in newspapers (DeWeese, 1977). A few years later, LexisNexis began offering digital editions of American newspapers that were accessible electronically through remote access terminals. The 1980s saw a resurgence of interest in content analysis, particularly in communication and political science, driven by the development of personal computers and the growing availability of digital and digitized media content. Simultaneously, with advances in artificial intelligence research, approaches beyond dictionary classification and co-occurrence analysis were rediscovered. However, the initial enthusiasm (e.g., Weber, 1984, p. 142) was tempered by the realization that computers could not *understand* texts in the foreseeable future (van Cuilenburg, Kleinnijenhuis, & Ridder, 1988). Nevertheless, social scientists, most notably Dutch researchers from the CETA project, began to investigate the possibilities of syntactic-semantic content analysis with computer assistance, aiming to move beyond word counts and purely statistical approaches to text analysis.

The proliferation of the Internet since the 1990s has resolved the issue of limited access to machine-readable text content. Instead, the current challenge lies in managing the increasing volume and diversity of textual (and audio-visual) content produced and shared online. The 2000s and 2010s saw increased computational power of personal computers and the introduction of more user-friendly software packages. These developments made CTA accessible to a broader range of applied social scientists from communication and related disciplines (Boumans & Trilling, 2016; Grimmer & Stewart, 2013; Hase et al., 2023; Macanovic, 2022). While dictionaries and simple word co-occurrence analyses remained popular, they were supplemented by more advanced techniques such as supervised machine learning or topic models. However, these techniques still relied on a bag-of-words representation of texts, which largely ignored the syntactic structure of texts.

This began to change in the 2010s and 2020s when deep learning approaches revolutionized many areas of natural language processing and computational text analysis. Neural network architectures, including word embeddings, recurrent neural networks, and transformers, enabled the learning of rich text representations that capture semantic and syntactic relationships. Embedding models gained popularity for representing words, sentences, and documents as dense

vectors that encode their meaning and context. The availability of large pre-trained language models through open-source platforms and closed-source but user-friendly programming interfaces facilitated powerful transfer learning approaches. These approaches involve fine-tuning the models on specific downstream tasks using smaller labeled datasets or even zero-shot learning (i.e., new tasks without training data) (Laurer et al., 2024a; Törnberg, 2023, 2024). Social scientists increasingly adopted analysis techniques built with such models while critically examining bias, fairness, and interpretability (Kroon et al., 2024).

3. Representing text as data

While obtaining and processing machine-readable texts was historically challenging, today, it is primarily a matter of technical access, ethical considerations, and legal restrictions. The critical methodological issue is structuring and condensing the texts to make them amenable to CTA. This section introduces three common text representations: bag of words, semantic networks, and embeddings.

3.1 Texts as bags of words

The bag-of-words (BoW) representation is a simple yet effective technique for representing text data in natural language processing tasks. In the BoW approach, a text (such as a sentence or a document) is represented as an unordered collection, or "bag," of its constituent tokens, disregarding grammar and word order but keeping track of token frequencies (Grimmer & Stewart, 2013).

The BoW representation requires several text preprocessing steps to clean and normalize the text data (Manning & Schütze, 1999). Commonly used techniques include tokenization, lowercasing, removing punctuation and numbers, stemming, lemmatization, and removing stop words. While text preprocessing is essential for model performance when analyzing BoW representations, different preprocessing choices can substantially impact the analysis results. This opens up researchers' degrees of freedom and calls reproducibility into question (Denny & Spirling, 2018; Maier et al., 2020; Pipal et al., 2023). After tokenization and the other preprocessing steps, each unique token in the corpus becomes a feature in the resulting document-term matrix, where each row represents a document, each column corresponds to a token, and each cell contains the frequency or occurrence of a token in a document.

While using individual words as tokens is most common, the BoW model can be extended to incorporate n-grams - contiguous sequences of n items from a text sample. N-grams allow capturing some local word order and context. For example, with bigrams (n=2), the phrases "not good" and "very good" would be treated as distinct features, while the token "good" would be indistinguishable under a unigram (n=1) model. However, using higher-order n-grams can quickly lead to high-dimensional and sparse representations. Besides using n-grams, the BoW model makes the simplifying assumption that the order of tokens in a document can be neglected and that the content of documents is represented as a multiset of their tokens. While this representation loses the original word order, syntactic structures, and word-level context, it still preserves essential information about the content and topics of a document.

The main advantages of the BoW approach are its simplicity, efficiency, and ability to extract numeric features from text that can be used as input to statistical machine learning models. BoW models are straightforward to understand and implement, computationally efficient, and can handle large amounts of text data. However, the approach also has several

limitations. One major disadvantage is that BoW is unsuited for short texts like social media messages or headlines. Such texts often lack sufficient word co-occurrence data and context for BoW to capture meaningful patterns. Social media content tends to be noisier, using more abbreviations, slang, and irregular structures that confuse BoW models. The BoW representation of short texts is typically very sparse, as most words in the vocabulary are absent. This sparsity can hurt the performance of machine learning models trained on BoW features. Other general limitations of BoW include losing word order information, ignoring word meaning and semantics, and struggling with frequent or rare words. Despite its drawbacks, due to its simplicity and efficiency, the BoW model remains a popular baseline and starting point for many CTA applications.

3.2 Texts as semantic networks

Semantic network analysis is a method for representing and analyzing the meaning and relationships between concepts in a text corpus (van Atteveldt, 2008; van Cuilenburg et al., 1988). In semantic networks, nodes represent semantic concepts such as words, phrases, or topics, while edges represent associations or relationships between those concepts, including co-occurrence, similarity, or causality.

A common approach to constructing semantic networks from text is identifying the co-occurrence of terms within a specific context window, such as a sentence, paragraph, or document. The more frequently two terms appear together, the stronger their association in the resulting network. Automatic extraction of key concepts and relationships from text to build the semantic network can be achieved using statistical and natural language processing techniques, such as parsing linguistic dependencies. Compared to the BoW representation, semantic networks retain more information beyond single tokens. The relational structure of the data necessitates more complex data representations than the rectangular document-feature matrix, often using nested tree structures.

Representing text as a semantic network can uncover implicit patterns and structures of meaning that may not be apparent from directly reading the text. Visual analysis of the network structure, such as identifying densely interconnected clusters of related terms, can reveal the text's main themes, frames, and associations. Additionally, quantitative network analysis measures can be applied to compare the centrality of different concepts and the strength of their relationships.

The network representation enables visual and quantitative analysis of semantic patterns in text, providing insights complementary to other text analysis methods. However, its adoption in applied social science research has remained limited. Initially, the much less complex BoW representations could increasingly rely on larger data and more powerful statistical models, compensating for some of their deficits compared to semantic networks. Subsequently, the more user-friendly embedding approaches, which share some of the advantages of semantic networks without requiring a deeper linguistic understanding of the texts, superseded the use of semantic networks in applications that required contextual information.

3.3 Texts as dense vectors (embeddings)

Word embeddings have revolutionized computational text analysis by enabling the representation of words as dense vectors in a high-dimensional space. These vector representations capture semantic and syntactic relationships between words, allowing machines

to understand and process natural language more effectively (Kroon et al., 2024). Early word embedding models, such as *word2vec* (Mikolov et al., 2013) and *GloVe* (Pennington et al., 2014), learn word vectors based on the distributional hypothesis, which states that words occurring in similar contexts tend to have similar meanings. For example, *word2vec* uses a neural network to predict a target word given its surrounding context words or vice versa. *GloVe*, on the other hand, learns word vectors by factorizing a word co-occurrence matrix. These models have been widely adopted because they capture meaningful word relationships.

While word embeddings have proven highly effective, they have limitations. Each word is represented by a single vector, regardless of its context, which can be problematic for polysemous words with multiple meanings. Moreover, word embeddings do not capture higher-level semantic information in sentences or documents. Recent advances in transformer-based models, such as *BERT* (Devlin et al., 2019) and *GPT* (Radford et al., n.d.), have introduced contextualized word embeddings to address these issues. These models use an attention mechanism to learn dynamic word representations that adapt to the surrounding context, enabling them to disambiguate word senses and capture more nuanced semantic relationships.

Building upon the success of contextualized word embeddings, researchers have developed methods to generate sentence and document embeddings. These embeddings aim to encapsulate the overall meaning of a piece of text into a single vector representation. Approaches such as *doc2vec* (Le & Mikolov, 2014) extend the *word2vec* framework to learn document embeddings, while others leverage pre-trained transformer models to generate sentence embeddings by pooling or averaging the contextualized word embeddings. These higher-level embeddings enable document classification, clustering, and semantic similarity analysis tasks.

One key advantage of embedding representations is that they often require only minimal preprocessing of the texts. For example, stemming and lemmatization become unnecessary when words are projected to the vector space where different word forms are very close but not identical. This retains the information that they are very similar features, compared to a BoW document-term matrix where two word forms are treated as entirely different, unrelated features. Another advantage is the availability of multilingual and even multimodal embeddings. The former allows for analyses across language boundaries without the intermediate translation step (Licht, 2023). While still in earlier stages of development, the latter promises to enable an integrated content analysis of texts, visuals, and audio signals (Li et al., 2019). Importantly, embedding models are well-suited for shorter texts like social media posts. Traditional bag-of-words approaches struggle with the sparsity and lack of context in short texts, while word and sentence embeddings can effectively capture semantic similarities even with limited information (Kroon et al., 2024). However, these benefits come at the potential cost of using pre-trained models, which might contain biases present in the training data that can propagate to downstream applications (Bender et al., 2021).

4. A typology of CTA approaches

Similar to general statistical learning methods, approaches to CTA can be divided into unsupervised and supervised methods. Unsupervised methods aim to develop a function to classify messages into *a priori* unknown categories or along unknown dimensions. In contrast, supervised approaches assign messages to pre-defined categories or dimensions (Grimmer & Stewart, 2013). A typical example of an unsupervised statistical method is cluster analysis, in which similar objects are grouped automatically. Unsupervised approaches have the advantage of

being fully automatic. A researcher can define the relevant features or clustering rules but cannot directly influence the outcome of the analysis. This property makes unsupervised methods attractive for descriptive or explorative studies but unsuitable for hypothesis-driven content analyses. While a fully automated analysis can provide results quickly and without much prior effort, making sense of the results can be complex and leaves much room for subjective interpretations.

A typical example of a supervised statistical method is logistic regression, in which cases are predicted to be in one of two groups based on a function of predictor variables. Supervised approaches to CTA require human intervention in the classification process. Researchers must provide either pre-defined classification rules or classified example texts from which classification rules are derived. Because the researcher can define the outcomes of supervised approaches, they produce results that are straightforward to interpret. However, since they require human judgment, they are more costly in terms of time and effort and are subject to human errors in the classification or definition of rules and instructions. The errors can, in turn, bias the results of an analysis based on the classification (Bachl & Scharkow, 2017).

4.1 Unsupervised approaches

4.1.1 Text statistics

CTA based on text statistics typically builds on the BoW representation of texts. The techniques assume that inferences can be made about the context of messages, their authors, and their reception from the frequency of words and n-grams. Because computers are demonstrably faster and more reliable in counting words, they have been in use since the introduction of CTA. Text statistics are frequently employed in stylometry and authorship research, as an author's relatively clear fingerprint can be generated by examining the frequency distributions of certain words (Mosteller & Wallace, 1964). Technically, this can be accomplished by simply summarizing the columns of one or more document-term matrices.

Similarly, selecting key terms from documents by comparing their frequency within a document and in a larger corpus is possible. The ratio of the term frequency and the inverse document frequency (TF/IDF, see Manning & Schütze, 1999) is commonly used to measure the relative importance of a term and used as a feature weight in subsequent analyses. Word clouds, a figure that displays the most essential words in a collection of texts sized by their relevance, are popular but controversial visualizations of a simple text statistic.

Text statistics is also used to determine the formal readability and comprehensibility of texts, which assumes that specific text indicators can predict the content's complexity and comprehensibility. Typical indicators are average word and sentence length, vocabulary diversity, and the frequency of punctuation marks. These text statistics can be used as properties of message origins or production, for example, to distinguish between content from broadsheet and tabloid newspapers, tailor content to the expected recipients, e.g., children or adults, or check the comprehensibility of a message. For example, Kleinnijenhuis (1991) used statistical readability measures of newspaper coverage to indicate their complexity and linked these measures to the readers' knowledge. Thoms and colleagues (2020) investigated whether corporate reports differed in readability depending on whether they communicated good or bad news.

4.1.2 Co-occurrence analysis

In unsupervised analyses of texts, both the frequency of individual words and their associations—the co-occurrence of specific features within messages—are of interest. Co-occurrence analysis is the bivariate extension of the word frequency analysis. It is based on the assumption that cognitively or semantically related constructs are also spatially close to each other. By looking at words within a pre-specified unit of text, such as documents, paragraphs, or sentences, the association (collocation) of specific terms can be summarized in a contingency table or a similarity matrix. For example, a document-term matrix can be multiplied with its transposed form to yield a term-term matrix. This, in turn, can be subjected to cluster analysis or multidimensional scaling, which can be used to condense and visualize word associations. An alternative approach, pioneered by Iker and Harway (1965) with their *Words* program, uses exploratory factor analysis directly on the columns of a DTM. Co-occurrence analysis has been used extensively in communication research to understand semantic relations between concepts and how they change over time. It has also been used in the meta-analysis of communication research: Doerfel and Barnett (1999) analyzed ICA conference paper titles using the popular *CATPAC* software (Doerfel & Barnett, 1996) to reveal divisional and topical associations.

Co-occurrence analysis is frequently used as a dimensionality reduction technique. Instead of working with individual columns of a document-term matrix, researchers can use latent semantic indexing to reduce the number of variables by summarizing co-occurring terms in larger components or indices, which are then used in subsequent analyses (Deerwester et al., 1990). While simpler forms of co-occurrence analysis have fallen out of fashion with the availability of more powerful models, it is noteworthy that the concept of co-occurrence remains integral to embedding-based models. The embedding vectors represent co-occurrence patterns in the training material.

4.1.3 Text clustering, text scaling, and topic models

Grouping texts by their content and, thereby, exploring the content of a large collection of texts is one of the most popular applications of CTA. The most straightforward approach is document clustering based on the document-term matrix (Grimmer & Stewart, 2013). Similar to co-occurrence analyses, document clustering assumes that messages containing the same features are semantically or thematically similar. To determine the similarity between two documents, one can compute the similarity or distance measure based on the feature vector of each document. The cosine or the Jaccard coefficient is often used since these are relatively independent of the text length, i.e., the number of relevant features (Manning & Schütze, 1999). The resulting document-document distance matrix is used as a starting point for various cluster-based analytical methods. As with co-occurrence analysis, only the number of clusters can be determined before the analysis (e.g., using k-means clustering) or retrospectively (for hierarchical agglomerative methods).

Document scaling methods build on a similar idea. Instead of assigning a text to one discrete group, these techniques analyze and position documents along a latent dimension, often representing political ideology or sentiment. *Wordscore* (Laver et al., 2003) and *Wordfish* (Slapin & Proksch, 2008) are popular text scaling methods. *Wordfish* is a completely unsupervised scaling model that uses word frequencies to estimate the positions of documents on a single dimension, assuming the frequencies follow a Poisson distribution. It does not require reference texts and simultaneously learns document positions and word weights. In contrast, *Wordscore*

requires some supervision, namely a set of two reference texts with known positions relative to each other, to estimate the positions of virgin texts. Both methods assume that word usage reflects the underlying positions of documents. A similar approach that goes beyond the estimation of political positions is latent semantic scaling (Watanabe, 2021). It aims to place texts on a latent semantic dimension by analyzing semantic proximity between some seed words provided by the user and other words in the corpus. These methods identify latent traits that characterize and differentiate the documents.

Topic models are a conceptual extension of the document clustering technique. The different varieties of topic models are arguably among the most popular CTA methods in communication (see Chen et al., 2023, for a recent review). In contrast to traditional clustering techniques, topic models are mixed-membership models. They assume that texts belong to several latent topics at once and that terms and their co-occurrences have different probabilities conditional on the topic. Therefore, topic models can be considered a combination of term and document clustering. Latent Dirichlet allocation (Blei, 2012) was the first type of topic model widely used in communication (Chen et al., 2023). Another popular choice are structural topic models, which can estimate the relationships of topic frequency with covariates, making it easy to analyze trends over time or differences between communicators or media outlets (Roberts et al., 2019).

There is an active methodological research stream in communication on applying and improving topic models. Maier and colleagues (2018) presented an early guideline on how to apply topic modeling and report the results. A recent review by Chen and colleagues (2023) discusses the use of the technique in communication. They critically conclude that, while topic models contributed substantially to exploratory research, many applications remained undertheorized and would benefit from more systematic validation (see also Bernhard et al., 2023). Several studies have investigated topic modeling of multilingual text collections (Chan et al., 2020; Lind et al., 2022; Maier et al., 2022). Other papers made recommendations on how topic modeling can be integrated into a larger content-analytical framework that includes both CTA and human input (Baden et al., 2020; Drohr & Ophir, 2019; Rinke et al., 2022).

While most of the published applications of topic modeling and related techniques were built on BoW representations, modern implementations take word or text embedding vectors as their input data. Text scaling methods can use word embeddings to estimate political ideology (Rheault & Cochrane, 2020). The BERTopic implementation (Grootendorst, 2022) starts with embedding the texts in a high-dimensional vector space. The dimensionality is reduced, and text clusters are identified in the low-dimensional space. Only after the clustering are the tokens of all texts in one cluster merged and analyzed in a bag-of-word logic. Embedding-based topic models have great promise for communication research. Many studies aim to explore the content of social media messages, which are typically too short for BoW-based clustering (Chen et al., 2023). Simon and colleagues (2023) showcased an early adoption of BERTopic in their large-scale analysis of information flows within the Dutch Telegramsphere.

Topic models and other text clustering or scaling techniques are expected to remain highly popular in communication research. These methods enable researchers to efficiently explore, describe, and, to a certain extent, *understand* previously unknown properties of large text collections. From a methodological standpoint, novel models that move beyond the BoW representation will address many limitations identified in previous applications. However, to

enhance the contribution of such studies to communication research, a more robust reflection on the results from a theoretical and conceptual perspective is still necessary.

4.2 Supervised approaches

4.2.1 Rule-based CTA: Dictionaries

For decades, dictionaries have often been used synonymously with CTA. The basic principles underlying this approach have barely changed since Stone et al. (1966) introduced the *General Inquirer*. Researchers develop a category system in which individual words (or other features) are defined to serve as indicators for the category of interest. The word list, or dictionary, must be constructed to be both exhaustive (all relevant features are scored) and specific (only relevant features are scored so that the risk of false positive classifications is minimized). This simplifies the analysis of term-document matrix problems. In every row, the number of relevant features from the dictionary is counted, and the documents are classified according to a threshold criterion. This approach enables researchers to quickly and reliably assign many texts to pre-defined categories. Because of the fully deterministic matching process, dictionary-based classification is perfectly reliable (in the sense that all classifications are fully reproducible); however, there is no room for fuzzy categories, double meanings, or contextual factors inherent to natural language. Accordingly, most of the earlier research on dictionary classification focused on text and term pre-processing to reduce language ambiguity and develop valid dictionaries for various research questions.

Dictionary-based classifications were considered attractive because of the quick and reliable classification process and because they promised reusable dictionaries that could foster collaboration and replication. However, this promise was only partially fulfilled. While several general-purpose dictionaries such as the *Harvard IV Dictionary*, the *Lasswell Values Dictionary*, or the *Linguistic Inquiry and Word Count* have been frequently applied, most dictionary-based studies used ad-hoc instruments that were rarely shared or reused. In addition, scholars have argued against the validity of dictionary classification because many theoretically relevant concepts that are relatively easy for humans to grasp cannot help build a reliable and valid dictionary despite extensive work. For example, Chan and colleagues (2021) demonstrated how using different dictionaries developed to measure the same construct can lead to completely different results. Moreover, when manual pre-processing is necessary to deal with spelling errors, homonyms, and other sources of measurement errors, dictionary-based content analysis requires more resources and effort than a traditional manual approach unless enormous quantities of messages have to be classified.

Despite these limitations, dictionaries are still actively used and developed, not least because they are so straightforward to understand and apply. This sets the method apart from machine learning methods, which require more technical skills from the researchers and perform the classification with algorithms that are harder to understand than matches with predefined indicators. There is ongoing methodological research into improving the application and development of dictionaries. A critical issue for comparative research is the creation of multilingual dictionaries (e.g., Lind et al., 2019). Researchers have recommended combining dictionaries with machine-learning approaches to improve the quality of the dictionaries (King et al., 2017) and to transfer dictionaries to new domains (Dobbrick et al., 2022).

Embedding-based models have also impacted dictionary-related work in several ways. For one, embeddings can be used to construct dictionaries or extend them beyond some seed

terms (Amsler, 2020; Liang et al., 2023; Stoll et al., 2023; Widmann & Wich, 2023). The researchers compile an initial set of terms from the literature to measure the construct. They then expand the dictionary by the nearest neighbors of the seed terms in an embedding space. A more complete representation of the construct of interest is achieved by reiterating this procedure and validating the expanded dictionaries against standard data. The new dictionary is then matched against the texts as usual for classification. The advantage of this approach is that the result is still a dictionary that can be easily understood, modified, and applied. The disadvantage is that, as a dictionary, it still relies on exact string matching for classification with all its shortcomings and requirements on preprocessing.

Another approach is based on training a word embedding model on the texts of interest and then computing the distances of target words and dictionary entries in the embedding space. For example, Andrich and colleagues (2023) trained embedding models on US American news articles. They then computed the distances between politicians and a dictionary of adjectives indicative of ten trait groups to measure gender-stereotypical representations. Similarly, Kroon and colleagues (2021) used the distances between ethnic categories and low- and high-thread terms in embedding representations of Dutch news to study stereotypical depictions of in- and outgroup minorities. The primary advantage of measuring distances in custom embedding spaces is that continuous distances do not require exact matches and can accommodate different word forms or misspellings. Furthermore, these distances carry more information than a binary match/mismatch indicator. However, a significant disadvantage is that this approach is only feasible with large collections of texts and considerable computational resources necessary for training custom embedding models. Additionally, the distance metric is less intuitive than the straightforward "number of matches" measure.

"Distributed Dictionary Representations" (Garten et al., 2018) is a related approach that can rely on publicly available embedding models trained on extensive text collections or custom embedding models. Like traditional dictionary analysis, a list of terms is compiled to represent a construct. However, instead of matching the dictionary with BoW representations of the texts, the similarity between the construct and a text is measured by the distance in the embedding space. Each text receives a score from -1 (entirely dissimilar) to +1 (identical). For example, Thiele (2024) applied this method to measure populism in comments to COVID-related social media posts by news organizations. Like measuring distances between target words and dictionary terms in the space of custom embedding models, this approach has the advantages of not requiring discrete string matching and providing more granular indicators, but the distance measure is less intuitive. When used with a pre-trained model, distributed dictionary representations can be applied to smaller text samples using user-grade compute resources, making it feasible for a broader range of applications. However, the choice of embedding model can substantially impact classification performance (Garten et al., 2018), and unknown biases in the training data can have downstream consequences. When used with a custom model, the previously described requirements of large samples and computing resources also apply.

Dictionary methods remain among the most frequently used CTA techniques despite—and because of—their simplicity. However, transitioning from string matching to measuring distances in embedding spaces has the potential to address some of the weaknesses associated with traditional dictionaries. These approaches share certain characteristics with few-shot and zero-shot machine learning models, described in the section after the next.

4.2.2 Supervised machine learning

A fundamental disadvantage of the dictionary method is that operationalization and classification differ significantly from how one would describe the construct of interest to another human, e.g., the annotators in traditional content analysis. Consequently, transferring expertise, previous results, or even instruments from manual content analyses to dictionary approaches is rarely possible. Supervised machine learning promises to make this possible by using messages classified by humans as training material for statistical learning algorithms, which in turn are used to classify large amounts of documents (Grimmer & Stewart, 2013). In supervised learning, training the computer is done by example: Instead of providing an extensive and exhaustive list of dictionary terms, the researcher provides example documents that belong to one category or another. The learning algorithms derive the classification rules through repeated examples and feedback.

The general workflow of CTA using supervised machine learning remains consistent regardless of the materials or topic being investigated. First, researchers define the constructs under investigation for human annotators, who then classify the training and test texts. Next, the texts are transformed into one or more data representations (see section "Representing texts as data"). One or more models are then trained on the training texts and validated on the test texts. These models can be relatively simple and computationally efficient, such as Naive Bayes, logistic regression, support vector machines, or more advanced, such as different shallow or deep neural network implementations. Based on their performance, one or a set of data representation and model combinations are selected if their quality is deemed sufficient. If none of the model-data combinations perform satisfactorily, parts of the process are adapted, and the subsequent steps repeated. Additional tests are conducted on new test data. The process is repeated until the desired model performance is achieved.

Supervised machine-learning approaches to CTA based on BoW models and simpler algorithms have shown mixed results in early evaluations in the social sciences (Grimmer & Stewart, 2013). Scharkow (2013) demonstrated that machine learning can reliably detect sports or politics in news articles. Similarly, Burscher and colleagues (2014) successfully classified the manifest indicators of relatively simple, holistic frames. In the political domain, Hillard and colleagues (2008) developed models that accurately predicted the topics of congressional bills. However, Scharkow (2013) found that more complex categories, such as the news factor controversy, were considerably more challenging to classify automatically. Although these findings were methodologically promising, the use of supervised machine-learning approaches to CTA in applied communication research remained limited. One likely reason was that theoretically more interesting concepts were still beyond the capabilities of these earlier implementations. Another reason seemed to be the complexities of the pre-processing pipelines of the BoW approach, which had the potential to bias the results in unexpected directions.

Embedding representations of texts and more advanced machine-learning algorithms have helped to address these challenges. Kroon and colleagues (2022) compared the impact of bag-of-words (BoW) and word-embedding representations of Dutch news articles. Still using relatively simple statistical algorithms, such as support vector machines and stochastic gradient descent, they demonstrated that models trained on word embeddings required less training data and achieved better performance in classifying policy issues and frames compared to the same models trained on BoW data and dictionaries. Rudkowsky and colleagues (2018) showed the superior performance of a neural network classifier based on an embedding representation for sentiment

analysis of sentences from parliamentary speeches compared to a multinomial Naive Bayes classifier based on a BoW representation. Van Atteveldt and colleagues (2021) reported similar results for measuring economic sentiment in news headlines, with a convolutional neural network based on word embeddings outperforming BoW-based supervised machine learning and dictionaries.

Large language models, deep-learning models based on the transformer architecture, have defined the next wave of innovation. These models have two main advantages compared to earlier methods. First, they can account for word order and context, allowing them to distinguish between identical terms that have different meanings depending on the context. Second, their multilayer architecture makes them ideal for fine-tuning. Pre-trained models can be adapted for new classification tasks instead of requiring completely new training. For example, Viehmann and colleagues (2023) showed that domain- and language-adapted BERT-based models performed best in detecting the stance of Twitter messages towards an issue. Using data from van Atteveldt and colleagues (2021), they also demonstrated the superior performance of adapted transformer models in classifying the sentiment of economic news. Similar results hold for classifying discrete emotions in political texts, with fine-tuned transformers considerably outperforming previous supervised machine-learning approaches based on word embeddings and dictionaries (Widmann & Wich, 2023). Laurer and colleagues (2024a) added natural language inference capabilities to transformer models by pre-training on classification tasks from various domains, making fine-tuning to new classification tasks more efficient. Their headline result in the political context shows that training their model on 500 labeled texts is about as effective as training older models on approximately 5,000 texts.

The modern transformer models are more performant and much easier to use than any previous machine learning setup for CTA. Advanced embeddings and the availability of adapted models and tokenizers for many domains and languages reduce the formerly tedious pre-processing tasks to a few reasonable default choices. Platforms like Hugging Face (https://huggingface.co/) are increasingly used to share pre-trained models for many tasks, which can be fine-tuned or used out of the box. Easy-to-use software makes the techniques available to the average social scientist by directly connecting to the platforms and using the models in established analysis pipelines (e.g., Chan, 2023; Kjell et al., 2023; Rajapakse, 2021).

4.2.3 Few-shot and zero-shot classification

Large language models have recently introduced few-shot and zero-shot approaches to classification tasks, bringing significant innovations to CTA. In few-shot learning, a model is trained on a minimal number of samples before classifying new texts, sometimes using as few as one example per class. Zero-shot approaches take this further by performing classification without any training data. Laurer and colleagues (2024a) presented natural language inference models capable of zero-shot classification on new tasks due to their training on diverse classification tasks. These models assign probabilities to each class based on a text and a short hypothesis about its classification (e.g., “This text is about {class},” where {class} is a list of classes). While their model can still be improved by task-relevant training data, the zero-shot classification beat the random and majority baselines in many of their test cases.

The public introduction of ChatGPT (https://chat.openai.com/) in late 2022 popularized generative language models, prompting many researchers to experiment with using these models for few-shot or zero-shot classification. In this approach, researchers provide a generative

language model with a prompt containing a category, its classes, definitions, possibly example text, and additional instructions. The model then classifies the provided text based on this information. It can also generate text to explain its "reasoning" for choosing the respective class (Törnberg, 2023). For example, Gilardi and colleagues (2023) tested the zero-shot performance of GPT-3.5, the model behind the first public version of ChatGPT, in several classification tasks typical for communication research, such as classifying frames, topics, and sentiment in tweets and newspaper articles. They concluded that the model outperformed crowd workers at a fraction of the cost and that model performance was sufficient compared to expert annotators for most tasks. Heseltine and Clemm von Hohenberg (2024) found that GPT-4, the successor to GPT-3.5, could reliably classify whether a tweet was political, negative, its sentiment, and its political ideology (left-wing, centrist, or right-wing). Moreover, the results were almost as good for tweets from Chile, Germany, and Italy and longer texts (news articles from U.S. news outlets) as for English tweets from U.S. politicians. Many similar results have been published as preprints over the past two years.

Few-shot and zero-shot approaches share similarities with dictionaries that distinguish them from supervised machine-learning approaches. The model that performs the classification task remains fixed during the study, akin to the term list and scoring rules of a dictionary approach. Rather than learning from training data by adapting model weights, these approaches rely solely on information from the instruction and texts for classification. As a result, the main focus for researchers becomes constructing and improving these inputs through an iterative process known as "prompt engineering" (Törnberg, 2024). Human-classified texts are only necessary for model validation when developing a new instruction or transferring an instruction to a new model or texts from a different population. By reducing the need for time- and cost-intensive human classifications, few-shot and zero-shot classification techniques can increase the accessibility and flexibility of large-scale CTA. However, the computing resources required for model inference with state-of-the-art generative language models exceed what is available to most social scientists, limiting individual researchers' flexibility. Additionally, commercial companies offer many of the most performant models as closed-source cloud services, posing challenges to open science principles such as reproducibility and raising concerns about biases encoded in the models (Bender et al., 2021; Spirling, 2023). Natural language inference models offer a computationally efficient alternative for classification tasks, as they do not generate text output and can run on modern end-user hardware. Moreover, most of these models are available as open-source software (Laurer et al., 2024b).

5. Outlook

Methods based on embeddings, transformers, and large language models are poised to replace traditional bag-of-words approaches sooner rather than later. This shift represents a positive development for the field, as these advanced techniques offer superior performance and robustness compared to their predecessors. By reducing the impact of researchers' choices in text preprocessing, these methods can help to standardize analyses and improve the reproducibility of findings. However, as we embrace this change, it is crucial to remain aware of the potential drawbacks associated with these cutting-edge approaches. One concern is the presence of unclear biases in the underlying models, which may stem from the training data used. Additionally, these methods often require more computational resources than traditional techniques, which could limit their accessibility to some researchers. Furthermore, reliance on closed-source models from commercial organizations raises questions about transparency and reproducibility.

To address these challenges, the communication research community should prioritize building upon open-source models and fostering interdisciplinary cooperation. By collaborating with experts in computer science, linguistics, and other relevant fields, we can develop robust, transparent, and accessible tools for CTA. This collaborative approach will enable us to leverage the strengths of CTA techniques while mitigating their potential weaknesses. Ultimately, by embracing the potential of embeddings, transformers, and large language models while also addressing their limitations, communication researchers can focus on their core expertise: improving measures in relation to theories and validating the performance of these measures. As we move forward, it is essential to balance harnessing the potential of cutting-edge computational methods and maintaining a strong connection to the fundamental principles of social science research.

References

Andrich, A., Bachl, M., & Domahidi, E. (2023). Goodbye, gender stereotypes? Trait attributions to politicians in 11 years of news coverage. *Journalism & Mass Communication Quarterly*, *100*(3), 473–497. https://doi.org/10/gsxh3w

Amsler, M. (2020). *Using lexical-semantic concepts for fine-grained classification in the embedding space* [Dissertation, University of Zurich]. https://doi.org/10.5167/uzh-189884

Bachl, M., & Scharkow, M. (2017). Correcting measurement error in content analysis. *Communication Methods and Measures, 11*(2), 87–104. https://doi.org/10/ghhzbn

Baden, C., Kligler-Vilenchik, N., & Yarchi, M. (2020). Hybrid content analysis: Toward a strategy for the theory-driven, computer-assisted classification of large text corpora. *Communication Methods and Measures, 14*(3), 165–183. https://doi.org/10/ghzn2b

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. https://doi.org/10/gh677h

Bernhard, J., Teuffenbach, M., & Boomgaarden, H. G. (2023). Topic model validation methods and their impact on model selection and evaluation. *Computational Communication Research, 5*(1). https://doi.org/10/gt2x55

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10/b39c

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

Chan, C. (2023). grafzahl: Fine-tuning Transformers for text data from within R. *Computational Communication Research, 5*(1). https://doi.org/10.5117/CCR2023.1.003.CHAN

Chan, C., Bajjalieh, J., Auvil, L., Wessler, H., Althaus, S., Welbers, K., Atteveldt, W. van, & Jungblut, M. (2021). Four best practices for measuring news sentiment using 'off-the-shelf'

dictionaries: A large-scale p-hacking experiment. *Computational Communication Research, 3*(1). https://doi.org/10/gt3d7x

Chan, C.-H., Zeng, J., Wessler, H., Jungblut, M., Welbers, K., Bajjalieh, J. W., van Atteveldt, W., & Althaus, S. L. (2020). Reproducible extraction of cross-lingual topics (rectr). *Communication Methods and Measures, 14*(4), 285–305. https://doi.org/10/gms9jj

Chen, Y., Peng, Z., Kim, S.-H., & Choi, C. W. (2023). What we can do and cannot do with topic modeling: A systematic review. *Communication Methods and Measures, 17*(2), 111–130. https://doi.org/10/gr4345

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407. https://doi.org/10/db4ft5

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis, 26*(2), 168–189. https://doi.org/10/gdjsqk

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://doi.org/10.48550/arXiv.1810.04805

DeWeese, L. (1977). Computer content analysis of "day-old" newspapers: A feasibility study. *Public Opinion Quarterly*, *41*(1), 91–94. https://doi.org/10.1086/268357

Dobbrick, T., Jakob, J., Chan, C.-H., & Wessler, H. (2022). Enhancing theory-informed dictionary approaches with "glass-box" machine learning: The case of integrative complexity in social media comments. *Communication Methods and Measures, 16*(4), 303–320. https://doi.org/10/gt28wz

Doerfel, M. L., & Barnett, G. A. (1996). The use of CATPAC for text analysis. *CAM Journal, 8*(2), 4–7. https://doi.org/10/dc8mgt

Doerfel, M. L., & Barnett, G. A. (1999). A semantic network analysis of the international communication association. *Human Communication Research*, *25*(4), 589–603. https://doi.org/10.1111/j.1468-2958.1999.tb00463.x

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior Research Methods*, 50(1), 344–361. https://doi.org/10/gc3z28

Gerbner, G., Holsti, O., Krippendorff, K., Paisley, W., & Stone, P. (1969). *The analysis of communication content*. Wiley.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30), e2305016120. https://doi.org/10/gsqx5m

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*. https://doi.org/10.48550/arXiv.2203.05794

Hase, V., Mahl, D., & Schäfer, M. S. (2023). The "computational turn": An "interdisciplinary turn"? A systematic review of text as data approaches in journalism studies. *Online Media and Global Communication, 2*(1), 122–143. https://doi.org/10/gt2kx9

Heseltine, M., & Clemm von Hohenberg, B. (2024). Large language models as a substitute for human experts in annotating political text. *Research & Politics, 11*(1), 20531680241236239. https://doi.org/10/gtkhqr

Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, *4*(4), 31–46. https://doi.org/10.1080/19331680801975367

Holsti, O. (1969). *Content analysis for the social sciences and humanities*. Addison-Wesley.

Iker, H. P., & Harway, N. I. (1965). A computer approach towards the analysis of content. *Behavioral Science*, *10*(2), 173–182. https://doi.org/10.1002/bs.3830100209

King, G., Lam, P., & Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science, 61*(4), 971–988. https://doi.org/10/gcgmwn

Kjell, O., Giorgi, S., & Schwartz, H. A. (2023). The text-package: An R-package for analyzing and visualizing human language using natural language processing and transformers. *Psychological Methods, 28*(6), 1478–1498. https://doi.org/10/gsmcq8

Kleinnijenhuis, J. (1991). Newspaper complexity and the knowledge gap. *European Journal of Communication, 6*(4), 499–522. https://doi.org/10/cmfpdq

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Sage.

Kroon, A. C., Meer, T. G. L. A. V. der, & Vliegenthart, R. (2022). Beyond counting words: Assessing performance of dictionaries, supervised machine learning, and embeddings in topic and frame classification. *Computational Communication Research, 4*(2). https://doi.org/10/gtw82b

Kroon, A. C., Trilling, D., & Raats, T. (2021). Guilty by association: Using word embeddings to measure ethnic stereotypes in news coverage. *Journalism & Mass Communication Quarterly*, 98(2), 451–477. https://doi.org/10/gt2xk5

Kroon, A., Welbers, K., Trilling, D., & van Atteveldt, W. (2024). Advancing automated content analysis for a new era of media effects research: The key role of transfer learning. *Communication Methods and Measures, 18*(2), 142–162. https://doi.org/10/gsv44t

Lasswell, H., Lerner, D., & Sola Pool, I. de. (1952). *The comparative study of symbols: An introduction*. Stanford: Stanford University Press.

Laurer, M., Atteveldt, W. van, Casas, A., & Welbers, K. (2024a). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis, 32*(1), 84–100. https://doi.org/10/gsgptm

Laurer, M., van Atteveldt, W., Casas, A., & Welbers, K. (2024b). Building efficient universal classifiers with natural language inference. *arXiv*. https://doi.org/10/m7cn

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review, 97*(2), 311–331. https://doi.org/10/ctpzr7

Liang, H., Ng, Y. M. M., & Tsang, N. L. T. (2023). Word embedding enrichment for dictionary construction: An example of incivility in Cantonese. *Computational Communication Research*, 5(1). https://doi.org/10/gt3d7s

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). VisualBERT: A simple and performant baseline for vision and language. *arXiv*. https://doi.org/10.48550/arXiv.1908.03557

Licht, H. (2023). Cross-lingual classification of political texts using multilingual sentence embeddings. *Political Analysis, 31*(3), 366–379. https://doi.org/10/gt2xk7

Lind, F., Eberl, J.-M., Eisele, O., Heidenreich, T., Galyga, S., & Boomgaarden, H. G. (2022). Building the bridge: Topic modeling for comparative research. *Communication Methods and Measures, 16*(2), 96–114. https://doi.org/10/gmxm6c

Lind, F., Eberl, J.-M., Heidenreich, T., & Boomgaarden, H. G. (2019). When the journey is as important as the goal: A roadmap to multilingual dictionary construction. *International Journal of Communication*, 13, https://ijoc.org/index.php/ijoc/article/view/10578.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning*, 1188–1196. https://proceedings.mlr.press/v32/le14.html

Macanovic, A. (2022). Text mining for social science – The state and the future of computational text analysis in sociology. *Social Science Research*, 108, 102784. https://doi.org/10/grhwwc

Maier, D., Baden, C., Stoltenberg, D., De Vries-Kedem, M., & Waldherr, A. (2022). Machine translation vs. multilingual dictionaries: Assessing two strategies for the topic modeling of multilingual text collections. *Communication Methods and Measures, 16*(1), 19–38. https://doi.org/10/gms9mp

Maier, D., Niekler, A., Wiedemann, G., & Stoltenberg, D. (2020). How document sampling and vocabulary pruning affect the results of topic models. *Computational Communication Research, 2*(2). https://doi.org/10/gt2x6g

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*. https://proceedings.neurips.cc/paper_files/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

Mosteller, F., & Wallace, D. (1964). *Inference and disputed authorship: The Federalist*. Addison-Wesley.

Osgood, C. (1959). The representational model and relevant research methods. In I. de S. Pool (Ed.), *Trends in content analysis* (pp. 33–88). University of Illinois Press.

Pipal, C., Song, H., & Boomgaarden, H. G. (2023). If you have choices, why not choose (and share) all of them? A multiverse approach to understanding news engagement on social media. *Digital Journalism, 11*(2), 255–275. https://doi.org/10/gt2w9r

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10/gfshwg

Popping, R. (2000). *Computer-assisted text analysis*. Sage.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (n.d.). *Improving language understanding by generative pre-training*. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Rajapakse, T. (2021). *Simple Transformers* [Software package]. https://simpletransformers.ai/

Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis, 28*(1), 112–133. https://doi.org/10/ghfjw4

Rinke, E. M., Dobbrick, T., Löb, C., Zirn, C., & Wessler, H. (2022). Expert-informed topic models for document set discovery. *Communication Methods and Measures, 16*(1), 39–58. https://doi.org/10/gmbkwd

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software, 91*(1), 1–40. https://doi.org/10/ggc8cz

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures, 12*(2–3), 140–157. https://doi.org/10/ghhzgh

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

Scharkow, M. (2017). Content analysis, automatic. In J. Matthes, C. S. Davis, & R. F. Potter (Eds.), *The International Encyclopedia of Communication Research Methods*. Wiley. https://doi.org/10.1002/9781118901731.iecrm0043

Shahin, S. (2016). When scale Meets depth: Integrating natural language processing and textual analysis for studying digital corpora. *Communication Methods and Measures*, 10(1), 28–50. https://doi.org/10.1080/19312458.2015.1118447

Simon, M., Welbers, K., C. Kroon, A., & Trilling, D. (2023). Linked in the dark: A network approach to understanding information flows within the Dutch Telegramsphere. *Information, Communication & Society, 26*(15), 3054–3078. https://doi.org/10/grhj9f

Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science, 52*(3), 705–722. https://doi.org/10/brh9q7

Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature, 616*(7957), 413–413. https://doi.org/10/gsqx6v

Stoll, A., Wilms, L., & Ziegele, M. (2023). Developing an incivility dictionary for German online discussions – a semi-automated approach combining human and artificial knowledge. *Communication Methods and Measures*, 17(2), 131–149. https://doi.org/10/gsnfdn

Stone, P. (1997). Thematic text analysis: New agendas for analyzing text content. In C. W. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (pp. 35–54). Lawrence Erlbaum Associates.

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

Thiele, D. (2024). How COVID-19 and the news shaped populism in Facebook comments in seven European countries: A computational analysis. *Computational Communication Research, 6*(1). https://doi.org/10/gt3dhm

Thoms, C., Degenhart, A., & Wohlgemuth, K. (2020). Is bad news difficult to read? A readability analysis of differently connoted passages in the annual reports of the 30 DAX companies. *Journal of Business and Technical Communication, 34*(2), 157–187. https://doi.org/10/gt3vpb

Törnberg, P. (2023). How to use LLMs for text analysis. *arXiv*. https://doi.org/mqx9

Törnberg, P. (2024). Best practices for text annotation with large language models. *arXiv*. https://doi.org/gtn9qf

van Atteveldt, W. H. (2008). *Semantic network analysis: Techniques for extracting, representing, and querying media content* [Dissertation, VU Amsterdam]. https://research.vu.nl/files/75843774/complete%20dissertation.pdf

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures, 15*(2), 121–140. https://doi.org/10/gh8dk3

van Cuilenburg, J. J. van, Kleinnijenhuis, J., & Ridder, J. A. de. (1988). Artificial intelligence and content analysis. *Quality and Quantity*, *22*(1), 65–97. https://doi.org/10.1007/bf00430638

Viehmann, C., Beck, T., Maurer, M., Quiring, O., & Gurevych, I. (2023). Investigating opinions on public policies in digital media: Setting up a supervised machine learning tool for stance classification. *Communication Methods and Measures, 17*(2), 150–184. https://doi.org/10/gsr7sv

Walter, D., & Ophir, Y. (2019). News frame analysis: An inductive mixed-method computational approach. *Communication Methods and Measures, 13*(4), 248–266. https://doi.org/10/ggjvnz

Watanabe, K. (2021). Latent semantic scaling: A semisupervised text analysis technique for new domains and languages. *Communication Methods and Measures, 15*(2), 81–102. https://doi.org/10/gmbkwh

Weber, R. P. (1984). Computer-aided content analysis: A short primer. *Qualitative sociology*, *7*(1), 126–147. doi:10.1007/bf00987112

Widmann, T., & Wich, M. (2023). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. *Political Analysis*, 31(4), 626–641. https://doi.org/10/gr9dpq

Züll, C., & Mohler, P. Ph. (2001). *Computerunterstützte Inhaltsanalyse: Codierung und Analyse von Antworten auf offene Fragen* [Computer-aided content analysis: classification and analysis of responses to open-ended survey questions]. Zentrum für Umfragen, Methoden und Analysen. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-201405