Content Analysis, Automatic

MICHAEL SCHARKOW
Zeppelin University, Germany

Automatic content analysis (ACA) is a technique for coding messages with the help of computer algorithms. Unlike computer-aided content analysis, ACA is defined as any method in which the actual coding decision, that is, assigning codes to documents or single textual or audiovisual elements, does not require human judgment and therefore is performed automatically. Since ACA relies on the computing capabilities of machines rather than human coders, it can be applied to very large documents. Moreover, automatic coding is highly reliable in that any analysis can be reproduced exactly given the same material and software and all errors are deterministic, that is, stemming from misspecification or program errors.

Historical development

The history and development of ACA can be understood through three central themes: (i) the concept of content analysis and its computational implementation, (ii) the development of software for automatic data processing and analysis, and (iii) the provision and analysis of digital or machine-readable documents. The first phase of development of computer-assisted and automatic content analysis in the late 1950s was mainly characterized by experiments with the computer. For this, social scientists had to first learn to program the computer, and in most cases, this involved the support of the science departments. The first few experiments were almost exclusively limited to producing text statistics, that is, counting words, a technique that had been applied in many disciplines such as political science or literature since the 1920s (Holsti, 1969; Stone, Dunphy, Smith, & Ogilvie, 1966). Conceptually, these simple automatic analyses fell well behind the advancements made in conventional content analysis by Lasswell, Lerner, and de Sola Pool (1952) or Osgood (1959). However, early content analysts were increasingly troubled by the costs of manual coding and had high expectations from automatic approaches, the development of which was viewed as central to the success of the method (Stone, 1997). At this time, computer-assisted content analysis was a high-risk research domain, plagued with numerous problems. On the one hand, primitive computing hardware and software allowed only small amounts of text to be analyzed with a few variables (Iker & Harway, 1965). On the other hand, because of the dearth of machine-readable documents, all units of analysis had to be digitized in a complex and error-prone process on punch cards. Thus, in the early 1960s, ACA was neither cost-effective—half an

The International Encyclopedia of Communication Research Methods. Jörg Matthes (General Editor), Christine S. Davis and Robert F. Potter (Associate Editors).

DOI: 10.1002/9781118901731.iecrm0043

hour of computing time cost as much as the monthly salary of a secretary (Stone, 1997, p. 42)—nor less laborious than manual coding.

The introduction of the General Inquirer (Stone et al., 1966) and Words (Iker & Harway, 1965) programs marked the first milestone in the development of ACA. Not only did these software packages enable automatic content analyses with relative ease, they also represented the first prototypes of two alternative approaches to automatic analysis—dictionary-based coding and co-occurrence analysis—that dominated the discipline for the next few decades. These innovative studies presented at the Annenberg Conference (Gerbner, Holsti, Krippendorff, Paisley, & Stone, 1969) defined the standards for ACA for years to come. While technological advancements continued in the 1970s, enabling more social scientists to use computers, the conceptual and methodological development of ACA slowed down, and interest in these techniques waned (Weber, 1984, p. 127), especially in the United States. Meanwhile, many German scholars followed the research program of the General Inquirer by developing dictionaries and software (TEXTPACK), primarily designed for the coding of open-ended survey questions but also used for the analysis of documents.

The unavailability of documents in machine-readable form was unresolved until the late 1970s; therefore, many studies were based only on archive material or non-news media. DeWeese (1977) was the first to collect daily media content automatically by using the increasingly common electronic typesetting machines used in newspapers. In 1979, the service provider LexisNexis started creating digital editions of American newspapers available electronically via remote access terminals. With the development of the personal computer and the further availability of digital(ized) media content, the 1980s witnessed a phase of renewed interest in ACA, especially in communication and political science. Fan (1988) demonstrated the potential of ACA in longitudinal studies of agenda setting and framing. A research team led by Schrodt developed software to automatically extract international events from the Reuters wire service messages, which later led to the program KEDS (Schrodt & Donald, 1990).

At the same time, with advances in artificial intelligence research, approaches beyond dictionary coding and co-occurrence analysis were rediscovered. The initial enthusiasm of Weber (1984, p. 142) and others, however, was dulled by the realization that computers would not be able to *understand* texts in the foreseeable future (van Cuilenburg, Kleinnijenhuis, & de 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, with the objective of moving beyond word counts and purely statistical approaches of text analysis. Despite the many advancements in computer linguistics, complete syntactic and semantic analyses of texts continue to be one of the biggest methodological challenges in ACA and an active area of interest for communication scholars (van Atteveldt, 2008).

Since the 1990s, and the spread of the Internet, the problem of limited availability of digital text content has been solved. In fact, the current challenge is to deal with the increasing amounts of textual and audiovisual information that is produced and distributed online. Given the increased computing capacity and the development of powerful statistical algorithms, many social scientists now rely on the bag-of-words approach, in which the syntactic structure of texts is largely disregarded. However, the

application of (supervised) machine learning, which draws on hand-coded training material rather than an extensive dictionary or rule set, has enabled scholars to combine traditional manual content analysis with automatic coding in a natural way (Scharkow, 2013).

In recent years, ACA—under different labels—has become highly popular, not only in computer science and related disciplines that focus on algorithmic and technical issues, but also in applied social research, mainly political science (Grimmer & Stewart, 2013) and communication (Boumans & Trilling, 2016). However, ACA has still not made its presence felt in mainstream social science research. Only a handful of pages are devoted to ACA in the current editions of textbooks on content analysis (Riffe, Lacy, & Fico, 2014). Moreover, much of the material focuses only on the dictionary and co-occurrence approaches from the 1960s and rarely discusses newer techniques such as unsupervised or supervised machine learning. This is also true of the two monographs that explicitly discuss ACA (Popping, 2000; Weber, 1984).

Research process

In principle, the research process in ACA is similar to traditional content analysis (Krippendorff, 2012), diverging only in those areas where human judgment is replaced with programmatically defined rules. ACA begins with the provision of machine-readable content. As detailed earlier, the digitization of media content has shifted the problem from data acquisition to sampling, storing, and managing potentially very large collections of texts and other subject matter. Once a piece of text from the Internet or other sources exists in machine-readable form, it must often be cleaned up for further processing, for example, by removing irrelevant or nontextual content. This is especially necessary when working with loosely structured data such as web pages or PDF documents, which often contain little content of interest and much *boilerplate content* such as navigation elements, advertising, or layout markup. In contrast to manual content analysis, where coders can simply be instructed to ignore or delete irrelevant content during the coding process, automatic filtering is not only necessary for any ACA—as the coding software cannot simple overlook content—but also difficult to implement with a deterministic rule set. In order to ensure long-term availability and reproducibility, messages are often converted to standardized formats such as Unicode text, XML, bitmap images, or video files.

In any content analysis in which the selection and coding units are not identical, the material must unitized before the actual coding. In manual analyses, this is mostly done either before or during the actual coding (Krippendorff, 2012, ch. 5). Typical examples are the segmentation of broadcast news into individual news segments or the breakdown of longer documents into smaller units such as claims. While manual unitization often relies on semantic structure, automatic approaches predominantly rely on formal criteria, that is, punctuation or paragraph marks. In practice, researchers often have to choose between using semantically rich constructs such as assertions, which require manual unitization, or working at the paragraph or sentence level, which is possible using fully automatic unitization procedures.

The next step in ACA is the definition of rules for feature selection, that is, deciding which properties of a message are considered relevant for the analysis. With textual data, this is relatively straightforward. In most cases, words (unigrams) or word sequences of a defined length (n-grams) are used. Features at the character or sentence level are easily analyzed but rarely used, as most text theories in communication research are based on words and their relations. There are, however, many examples for character-level analyses in computational linguistics, some of which could be used to augment traditional word-based content analyses. When using ready-made software packages, feature selection and extraction are often only implicit in the analysis, that is, users can rarely define non-word features. With audiovisual content, previous research has only focused on the textual level, for example, the use of transcripts of television news or descriptions of images, or simply asking human coders to verbalize the messages while coding. However, for ACA, a variety of possible and potentially relevant features on the visual and auditory level can be extracted: from individual pixels and audio signals to higher level structures and patterns (Lerch, 2012; Sonka, Hlavac, & Boyle, 2014).

Even with a small sample of survey units, the number extracted features can be very large, and the resulting data too noisy to analyze. In ACA, most messages are therefore preprocessed before the actual analysis. Preprocessing serves two different purposes: (i) it reduces the number of features (variables), and (ii) it makes the retained features as specific as possible, for example, by augmenting them with contextual information. The most common feature-reduction techniques for textual data are (i) conversion of all words to lowercase; (ii) removal of so-called *stop words*, that is, words that are either extremely frequent or infrequent in a corpus, as well as number, punctuation marks, and other special characters; (iii) lemmatization or algorithmic stemming. Lemmatization is a process by which different word forms (plural, inflections, etc.) are replaced by its basic form: the lemma. Since this technique requires a large dictionary for all word forms, it is highly language dependent. Stemming performs the same function—replacing different word forms with their basic form—but is based on a rule set rather than a dictionary. For example, many stemming algorithms for English remove suffixes such as -ing or -ed from verbs in order to obtain an (artificial) word stem. Much research in computational linguistics is devoted to the development of reliable preprocessing algorithms, and a number of different algorithms and software libraries are available for this task (van Atteveldt, 2008; Manning & Schütze, 1999). However, preprocessing can also lead to feature reduction, which is detrimental to the actual coding task. For example, removing common idioms or negations may lead to the deletion of key indicators for classification (Scharkow, 2013). Basically, if an analysis relies on the variance of features across messages, than a reduction in variance by means of preprocessing can have negative consequences.

The second type of preprocessing seeks to augment features with contextual information and aids the disambiguation of words, for example, distinguishes the adjective close from the verb close, or a goal in football from a goal in life. A frequently used technique for this purpose is part-of-speech tagging, by which all words are annotated according to their syntactic category, such as a noun or a verb. A related preprocessing step, called entity recognition, aims to detect specific proper names of persons, locations, or organizations in messages. A third disambiguation technique is anaphora resolution, which

Table 1 Document-term matrix.

	Government	*Secretary*	*Say*	*Key*	*Office*
Document1	1	0	1	0	1
Document2	0	1	1	0	0
Document3	0	0	0	1	1

involves replacing pronouns with the words they refer to. Many of these preprocessing steps are necessary when the ACA does not rely on the simple bag-of-words approach, but instead tries to extract semantic relations and meaning from sentences (van Atteveldt, 2008). While some preprocessing steps are relatively easy to implement (e.g., stop word removal or stemming), others require language- or domain-specific adjustments or manual annotation, which makes them unsuitable for large-scale applications.

Once the necessary preprocessing steps are applied in sequence, the actual analysis is often conducted using a document-term matrix (DTM, see Table 1), or more generally, an object-feature matrix, which is a reduced numerical representation of the content. In a DTM, documents constitute rows, terms constitute columns, and each cell indicates the occurrence of a term in a document. The DTM allows both simple analyses, such as summarizing the columns or rows, and more complex multivariate models, such as clustering by column or row. Many of these methods are described in the following section. The simple DTM structure is insufficient for analyses that are based on syntactic or semantic relationships. For these, the data is often stored in nested tree structures (van Atteveldt, 2008).

Typology of approaches to ACA

A number of different typologies for ACA have been suggested, either based on research questions or model criteria (see Grimmer and Stewart (2013) for a recent typology). On the basis of these suggestions, one can broadly distinguish between purely statistical bag-of-words models and linguistic models of texts (Schrodt & Donald, 1990, p. 4). The former assume that features (e.g., words) can be used as indicators of relevant constructs such as topics, frames, authorship, or text complexity. The latter assume that (only) relations between features can be used to extract meaning from messages, which implies that syntactic and semantic context must be considered in any analysis. Purely statistical approaches have the advantage of being largely independent of the coding material, as they only require a numerical object-feature matrix, whether that is based on textual or audiovisual, English or German texts, and so forth. In principle, newly developed algorithms and software programs can easily be plugged-in once the relevant preprocessing and feature extraction is complete. Statistical approaches therefore benefit from the abundant literature in the area of (statistical) machine learning. Linguistic approaches, on the other hand, are highly dependent on the material used (i.e., text) and the language features. Further, they require far more refined theories of text production and understanding. This makes it inherently more difficult for human

coders to process messages reliably, and even harder to automatize using computers (van Cuilenburg et al., 1988). Historically, statistical approaches have been the dominant paradigm in ACA, and even with the recent developments in computer linguistics, syntactic–semantic approaches are quite rare in applied communication research (van Atteveldt, 2008, is a notable exception).

According to the literature on machine learning, statistical approaches to ACA can be divided into unsupervised and supervised methods. Unsupervised methods aim to develop a function that can classify messages into a priori unknown categories, while supervised approaches assign messages to predefined categories or dimensions (Grimmer & Stewart, 2013). A typical unsupervised method from the social sciences is cluster analysis, in which similar objects are grouped together 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 makes unsupervised methods attractive for descriptive or explorative studies, but unsuitable for traditional hypothesis-driven content analyses. While a fully automated analysis can provide results very quickly and without much prior effort, the interpretation of the results can be difficult and highly subjective to much interpretation from the researcher. Supervised approaches require human intervention in the coding process, either by relying on predefined coding rules or by using manually coded example documents, from which coding rules are automatically derived. Because supervised approaches can be easily controlled by the researcher, 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 coding or rule definition.

Supervised approaches to ACA can also be distinguished by the way in which the automatic scoring or classification functions are derived. Either researchers explicitly develop coding rules which are then implemented in software (rule-based ACA), or they provide manually coded messages from which a machine learning algorithm derives the optimal coding function (example-based ACA, see Schrodt & Donald, 1990). Put differently, rule-based approaches are deductive, as they require researches to first develop a theory of text, including which features do and do not signify relevant categories, which is then used to develop coding rules. Example-based approaches are inductive: the researcher classifies messages according to a more or less loosely defined concept, and lets the computer determine which features best separate the different categories. Let us consider a simple thematic analysis in which a researcher wants to know if a message is related to a certain topic or not. Using the rule-based, deductive approach, a content analyst will try to compile a list of words, which indicate topic-related text. The occurrence of these words in a message is then taken as an indicator of category membership. In the example-based, inductive approach, the researcher develops a conventional codebook for the topic category and classifies a number of messages according to this codebook. These messages and the topic variable are then fed into a machine learning algorithm that simply tries to find the list of words that best separate the on-topic from the off-topic messages. Once this learning phase is accomplished, new messages are classified automatically, just as in the rule-based approach. Traditionally, supervised ACA has relied on rule-based techniques such as dictionary coding (see later).

Supervised scaling and text classification techniques have gained momentum only recently (Scharkow, 2013). It is important to note that barring the rule-development step, these approaches are very similar: both are based on the same (preprocessed) data, and the final classification of messages does not require human intervention.

Unsupervised approaches

Text statistics

ACA based on text statistics assumes 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, this technique has been in use since the introduction of ACA. Text statistics are frequently employed in stylometry and authorship research, as a relatively clear "fingerprint" of an author can be generated by examining the frequency distributions of certain words (Mosteller & Wallace, 1964). Technically, this can easily be accomplished by simply summarizing the columns of one or more document-term matrices. Similarly, it is possible to select key terms from documents by comparing their frequency within a document and in a larger corpus. 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.

Text statistics is also used to determine the readability and comprehensibility of texts, which assumes that certain text indicators can be used to predict the complexity and comprehensibility of the content. 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 aimed broadsheet and tabloid newspapers or to tailor content to the expected recipients, for example, children or adults, or to check the comprehensibility of a message (DuBay, 2004).

Co-occurrence analysis

In unsupervised analyses of texts, both the frequency of individual words and their associations—the co-occurrence of certain features within messages—are of interest. In principle, co-occurrence analysis is the bivariate extension of the word frequency analysis. Co-occurrence analysis is based on the assumption that cognitively or semantically related constructs are also spatially close to each other. By looking at words within a prespecified unit of text, such as documents, paragraphs, or sentences, the association (collocation) of certain terms can be summarized in a contingency table or a similarity matrix. For example, a DTM 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, mostly

to understand semantic relations between concepts and how they change over time. It has also been used in a content analysis of communication research: Doerfel and Barnett (1999) analyzed ICA conference paper titles using the popular CATPAC software to reveal divisional and topical associations. Co-occurrence analysis is frequently used as a simple dimensionality reduction technique. Instead of working with individual columns of a DTM, researchers often use a technique called latent semantic indexing (LSI) to reduce the number of variables by summarizing co-occurring terms in larger components or indices, which are then used in subsequent analyses (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990).

Like any exploratory factor or cluster analysis, co-occurrence analysis presents the problem that only the number of clusters or factors, not their meaning, is determined by the researchers, and the statistically derived factors or clusters are not always interpretable. Moreover, although the procedure itself is fully automatic, in many published co-occurrence analyses, a number of manual preprocessing steps were required, especially for the selection of relevant words. Nevertheless, co-occurrence analysis is a standard tool used by social science researchers, which is supported by many software packages.

Document clustering and topic models

The information contained in the document-term matrix can be used not only to analyze the association of single features but also to determine the similarities between coding units. It can be used for fully automatic categorization of documents, which is referred to as document clustering (Grimmer & Stewart, 2013). Similar to co-occurrence analyses, document clustering is based on the assumption that messages that contain the same features are semantically or thematically similar. To determine the similarity between two documents, one can compute a similarity or distance measure based on the feature vector of each document. The cosine or the Jaccard coefficients are often used for this purpose, since these are relatively independent of the text length, that is, the number of relevant features (Manning & Schütze, 1999). The resulting document–document distance matrix can then be used as a starting point for various cluster-based analytical methods. As with co-occurrence analysis, only the number of clusters can be determined, either prior to the analysis (e.g., using k-means clustering) or retrospectively (for hierarchical agglomerative methods). Moreover, if the outcome of interest is continuous rather than categorical, different dimensionality reduction techniques can be used in a similar fashion. This approach is called document scaling and used predominantly in political communication research (Grimmer & Stewart, 2013).

A conceptual extension of the classic document clustering technique has recently been introduced in communication research. Topic models are mixed-membership models which, in contrast to traditional document clustering, assume that messages belong to several latent topics at once, and that features as well as their co-occurrences have different probabilities conditional on the topic. Therefore, topic models can be thought of as a powerful combination of term and document clustering. Moreover, they can be extended to include additional contextual information such as topic hierarchies,

author information, or temporal ordering, which makes topic modeling highly attractive in many social science contexts (Grimmer & Stewart, 2013). In both document clustering and topic models, clusters or topics are automatically derived, so they must be not only interpreted with care, but possibly also validated retrospectively to ensure that the results are not biased by the sample composition or some parameter values.

Supervised approaches

Dictionary and rule-based coding

For decades, dictionary coding has often been used synonymously with ACA. The basic principles underlying this approach have barely changed since Stone et al. (1966) introduced the General Inquirer in the mid-1960s. Before the actual coding phase, 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 such that it is both exhaustive, that is, all relevant features are scored, and specific so that the risk of false positive classifications is minimized. This simplifies the analysis of term-document matrices. In every row, the number of relevant features from the dictionary are counted, and the documents are classified according to a threshold criterion. This approach enables researchers to quickly and reliably assign a very large number of documents to predefined categories. Because of the fully deterministic coding process, dictionary-based coding is perfectly reliable (in the sense that all codings are fully reproducible); however, there is no room for fuzzy categories, double meanings, or contextual factors that are inherent to natural language. Accordingly, most of the research on dictionary coding to date has focused on preprocessing, in order to reduce language ambiguity, and on developing valid dictionaries for various research questions. Dictionary-based coding was considered attractive not only because of the quick and reliable coding process, but also because it promised reusable dictionaries that could foster collaboration and replication. However, this promise was only partially fulfilled. While a number of general-purpose dictionaries such as the Harvard IV Dictionary, the Lasswell Values Dictionary, or the Linguistic Inquiry and Word Count (LIWC) (see Krippendorff, 2012, for an overview) have been frequently applied in the past decades, most dictionary-based coding uses ad hoc codebooks which are rarely shared or reused. In addition, scholars have argued against the validity of dictionary coding because many theoretically relevant concepts that are relatively easy for human coders to grasp cannot help build a reliable and valid dictionary, despite extensive work. Moreover, when manual preprocessing is necessary to deal with spelling errors, homonyms, and other sources of coding errors, dictionary-based content analysis can actually require more resources and effort than a traditional manual approach, unless really large quantities of messages have to be coded.

Closely related to dictionary coding are rule-based methods in which, besides the pure word level, syntactic or semantic information is used. Unlike purely thematic analyses, rule-based approaches are used to determine not only if a specific topic or an actor occurs but also how they relate to each other. Most rule-based methods use dictionaries, for example, lists of actors or locations, and then use this information in

combination with POS tagging or syntactic parsing for a detailed analysis. A typical rule-based coding task involves extracting all actions (marked by verbs) that connect nation states, for example, "Iraq invades Kuwait" or "Russia criticizes Israel." This approach was adopted by the KEDS/TABARI software for automatically extracting international event data from Reuters' headlines (Schrodt & Donald, 1990). A more general and potentially powerful technique for (semi-automatically) coding semantic relations is the CETA coding system (van Cuilenburg et al., 1988). However, despite the progress especially in syntactic parsing, most CETA-based analyses still require manual precoding, defining of exhaustive stakeholder and action lists, and domain-specific software modifications. Van Atteveldt (2008) illustrates both the great potential of the semantic network analysis—no other automatic method can be used to date for detailed frame or claim analysis—as well as the associated costs, especially considering a transfer of this approach to different subject areas or languages.

Machine learning

A fundamental disadvantage of the deductive, rule-based method is that the operationalization and coding differs significantly from traditional analysis. The process of developing a dictionary or implementing complex rules for a syntactic-semantic parser has little in common with traditional codebook development and coder training. Researchers who intend to use ACA often make the decision relatively early in the research design, so as to devote resources to developing dictionaries, preprocessing steps, and so on. However, it is rarely possible to transfer expertise, previous results, or even instruments from manual content analyses to automatic approaches. Supervised machine learning promises to make exactly this possible by using conventionally hand-coded messages as training material for a statistical learning algorithm, which in turn is used to code large amounts of documents (Sebastiani, 2002). In supervised learning, training the computer is similar to training a human coder: Instead of providing an extensive and exhaustive list of coding rules, the researcher provides example documents which belong to one category or another. Through repeated examples and feedback, coding rules are derived by the computer rather than predefined by the researcher. Compared to dictionary or rule-based coding, this approach has several advantages (Scharkow, 2013): (i) The procedure is independent of the coding material or topic as long as the training data are consistent. Machine learning algorithms use only the object-feature matrix, and therefore the same techniques can be used to classify text, images, audio, or video in any language and for any category system. The key step is the selection and extraction of relevant features. (ii) Little preprocessing and no manual adjustment of dictionaries or rules are necessary, so researchers familiar with traditional content analysis face fewer barriers. (iii) When manually coded digital content is already available, the machine learning algorithm can be trained and evaluated on these documents and codes with little additional effort. (iv) Since supervised classification is similar to traditional content analysis, the usual procedures for quality control, such as reliability and validity tests can be applied.

In recent years, machine learning approaches have been used with varying success in the social sciences (Grimmer & Stewart, 2013). It is important to note that while,

in principle, they can be used in all instances involving human coding decisions, in practice they may not work if the statistical learning algorithm cannot derive a good decision rule. This is often the case when categories rely on contextual knowledge of the coders and/or categories cannot be detected at the word level. For example, Scharkow (2013) showed that while machine learning can reliably be used to detect sports or politics articles, categories such as controversy are much harder to code automatically. As a rule of thumb, categories that are difficult for human coders to learn are less likely to be reliable when coded using supervised machine learning. Although many algorithms such as Naive Bayes, logistic regression, or support vector machines are available, in most applied communication research, the content categories and the selection of relevant features are far more important for the quality of the automatic coding than the choice of algorithm. In addition, using multiple algorithms in an ensemble is often more reliable than a single algorithm (Hillard, Purpura, & Wilkerson, 2008).

Benefits and limitations of ACA

In textbooks, the benefits of ACA are often discussed in terms of research economics and reliability. Automatic approaches are said to be cost-efficient, especially since the costs of coding additional messages are practically nil, and more reliable, since the computer does not tire, get bored, or commit random mistakes. The former argument is true only if one can (a) automate the complete research process, including unitization and preprocessing, and (b) if one uses pre-existing rules, dictionaries, or algorithms. If human intervention is required, the benefit disappears, as the costs of coding additional messages increase with the sample size. Moreover, previous experience has shown that developing reliable and valid rule-sets or dictionaries can require such extensive effort that it might be easier and cheaper to use traditional human coders. However, if it is necessary to code a large number of messages, then ACA is the only viable option. Beyond allowing for larger sample sizes, ACA has another methodological advantage: it enables researchers to test and apply different measures for the same construct with little additional cost. In most content analyses, messages are coded by a single coder using a single codebook, although measurement theory posits that multiple measures (preferably using different methods) will in general improve reliability and validity. Content analyses would therefore benefit from automatic approaches if they enable better measurement. Moreover, ACA is inherently more reproducible than the best manual analyses, since all the necessary steps are implemented in software, which by itself perfectly documents the research process.

The key limitation of ACA pertains to the validity of the codings. Given that computers cannot understand messages, how can researchers expect them to classify texts, images, or other subject matter according to defined criteria? The answer to this question contains two aspects. First, an instrument or technique cannot be valid or invalid—only its application and resulting inferences can be so described. If one is interested in the average length of a message, the presence of certain keywords, or identifying the most frequently co-occurring terms, then ACA is both more reliable and valid than manual content analyses. Of course, communication researchers are

not as interested in determining word length as in defining theoretical constructs that may be hard to formalize in computer code. Second, whether ACA can produce meaningful and valid results is an empirical question. In the past, efforts to create dictionaries for common categories such as negativity or rules that identify frames in news articles have failed, while other coding tasks have worked very well, even for seemingly subjective classification tasks such as email spam detection. Consequently, the pragmatic approach is to automate content analytic tasks that have been shown to work, and use human judgment where computers fail. Much can be learned from ACA even for manual content analysis. For example, the use of supervised text classification can help researchers refine the codebook by indicating vague or unclear coding instructions (Scharkow, 2013).

Perspective

Even though ACA is currently an emerging topic for communication scholars, many methodological aspects have received relatively little attention. Most importantly, the vast majority of ACA applications are focused on text, while images and audiovisual material are rarely coded automatically. Currently, few solutions exist for the automatic analysis of non-textual material, although face detection algorithms in image databases can be considered a starting point. Despite rapid developments in various fields of computer science, such as machine learning, computer vision, and computational linguistics, the adoption of new techniques in communication is rather slow, mainly because social scientists often lack the necessary technical expertise. From a methodological perspective, developments outside the sphere of social sciences are relevant because they often give rise to fundamental questions about content analysis. Since data-driven research in computer science or corporate environments is often very pragmatic when selecting relevant features of texts, pictures, or movies, a lot can be learned from this applied research, for example, that football match highlights can be identified only from background sounds, or that the number of repeated vowels in a word is an indicator of sentiment strength in online messaging. It remains a challenge for communication researchers to integrate these findings with previous knowledge about the uses and structure of media content.

In the era of ubiquitous online content, the analysis of both mass media and interpersonal messages is a challenge for communication scholars, not only because of the sheer volume of the data available but also of their nature (Shahin, 2016). Traditional content analysis, both manual and automatic, has focused mostly on relatively structured, static texts such as news articles, and categories such as issues, policy areas, or certain actors. However, most online content is structurally and thematically diverse and requires a complete rethinking of the content-analytic toolbox. The field of ACA will certainly see further improvement and adoption among communication scholars.

SEE ALSO: Big Data, Analysis of; Big Data, Collection of (Social Media, Harvesting); Code List/Codebook; Coding; Content Analysis, Quantitative

References

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. doi:10.1080/21670811.2015.1096598

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

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

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

DuBay, W. (2004). *The principles of readability*. Costa Mesa, CA: Impact Information.

Fan, D. (1988). *Predictions of public opinion from the mass media: Computer content analysis and mathematical modeling*. New York: Greenwood Publishing Group.

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

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. doi:10.1093/pan/mps028

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. doi:10.1080/19331680801975367

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

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

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: SAGE.

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

Lerch, A. (2012). *An introduction to audio content analysis: Applications in signal processing and music informatics*. Hoboken, NJ: John Wiley & Sons.

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

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

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

Popping, R. (2000). *Computer-assisted text analysis*. Thousand Oaks, CA: SAGE.

Riffe, D., Lacy, S., & Fico, F. (2014). *Analyzing media messages: Using quantitative content analysis in research*. New York: Routledge.

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

Schrodt, P., & Donald, C. (1990). Machine coding of events data. Paper presented at the annual meeting of the International Studies Association, Washington, DC, April.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. doi:10.1145/505282.505283

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. doi:10.1080/19312458.2015.1118447

Sonka, M., Hlavac, V., & Boyle, R. (2014). *Image processing, analysis, and machine vision*. Cengage Learning.

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* (pp. 35–54). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Van Atteveldt, W. (2008). *Semantic network analysis: Techniques for extracting, representing, and querying media content*. Charleston, SC: BookSurge.

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

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

Michael Scharkow (PhD, University of the Arts Berlin) is professor of communication at Zeppelin University, Germany. His research interests include empirical research methods, online communication, and media use.