Q Open, 2025, **5**, qoaf012
https://doi.org/10.1093/qopen/qoaf012
Advance access publication date: 13 March 2025
Article

Search term validation in agricultural economics: conceptual background and application

Richard Völker [1,*], Norbert Hirschauer [1], Fabienne Lind [2] and Sven Grüner [3]

[1]Institute of Agricultural and Nutritional Sciences, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
[2]Department of Communication, University of Vienna, Vienna, Austria
[3]Agricultural Economics, University of Rostock, Rostock, Germany

*Corresponding author. Martin Luther University Halle-Wittenberg, Institute of Agricultural and Nutritional Sciences, Agribusiness Management, Karl-Freiherr-von-Fritsch-Str. 4, D-06120 Halle (Saale).
E-mail: richard.voelker@landw.uni-halle.de

Received: September 9, 2024. **Accepted:** March 10, 2025

Abstract

Agricultural and environmental economists frequently use content analysis of textual data to gain a deeper understanding of public discourses that reflect the conflicting interests and attitudes of various stakeholders with regard to agricultural issues. These discourses encompass topics such as nitrogen leaching, climate change, biodiversity loss, and animal welfare. However, the procedural standards of content analysis established in communication studies are rarely fully adhered to due to a lack of interdisciplinary communication. This paper provides applied agricultural economists with a conceptual background to systematic search term validation to facilitate the transparent generation of high-quality databases for the content analysis of large datasets.

Keywords: content analysis; search terms; animal welfare; artificial intelligence; chatgpt; gemini.

JEL codes: A20, Q18

1. Introduction

Text as data is becoming increasingly important as a source of information (Gentzkow et al. 2019). Communication studies traditionally focus on how the media or specific media (e.g. particular newspapers) influence and reflect social discourses (Müller et al. 2018). Content analysis is the method of choice to do this (Lacy et al. 2015) and has been used, for example, in frame analysis (Matthes and Kohring 2008, Scheufele and Engelmann 2016, Bonfadelli and Friemel 2017). In recent years, agricultural economists and economists in general have begun to use media reports to gain a deeper understanding of the evaluations, beliefs, and attitudes of economic and other social actors. (cf. Baker et al. 2016, Müller et al. 2018, Hassan et al. 2019, Shiller 2019, Vecchio and Cavallo 2019, Jellason et al. 2022, Lehberger and Gruener 2023, Mizik 2023, Thompson et al. 2023, Schulze et al. 2024). The accessibility of digital databases, such as those provided by LexisNexis, Pro-Quest, or Google, in conjunction with computer-aided processing—also called automated content analysis (ACA) or distant reading—increasingly facilitates the analysis of extremely large text universes at relatively low cost (Grimmer and Stewart 2013, Boumans and Trilling 2016, Benoit 2020, Kroon et al. 2022).

Across various disciplines, search terms are the ubiquitous method to identify relevant textual data for content analysis (Mahl et al. 2022). A search term, search phrase, or search string (also referred to as a “chain of search terms”) is a tool created by a researcher to retrieve relevant textual data from a text universe (Stryker et al. 2006). As the design of search terms enables researchers to control the scope of the search and the database, it can substantially influence the results of the content analysis (Maier et al. 2020, Barberá et al. 2021). Defining valid search terms, that retrieve content which “truly represents the targeted concept of discourse” (Mahl et al. 2022: 4), raises two major challenges: first, they must ensure high “precision” and prevent meaningless data (noise) from entering the database; and second, they must ensure high “recall” and guarantee that as much of the relevant data as possible (ideally all of it) is included in the database. Meeting both challenges is difficult because there is a tradeoff: “narrow” search terms are likely to exclude most irrelevant textual data, but they are also likely to miss a fairly large share of the relevant data, and vice versa.

Stryker et al. (2006) conducted a review of content analyses published between 2000 and 2005, and found that fewer than 40 per cent of those using electronic databases provided information about the search terms used, and as few as 6 per cent discussed their validity. It appears little has changed. Many contemporary content-analytic studies (Velten et al. 2015, Janker and Mann 2020, Vindigni et al. 2021, Mohr and Höhler 2023) still do not report on how search terms were determined and which criteria were used to assess their validity (King et al. 2017, Riffe et al. 2019, Mahl et al. 2022). When there is a lack of transparency on how data were obtained for the analysis (“data generation”) reported results can be neither critically retraced nor replicated (M. Baker 2016, Nosek et al. 2022, Rommel et al. 2023) and their informational value remains opaque (Grimmer and Stewart, 2013, Lacy et al. 2015, Jünger and Geise 2022).

The lacking methodological transparency and rigor lead to the question of how researchers using ACA can efficiently obtain a high-quality database from large text universes and ensure that their analysis is based on as many relevant texts and as little noise as possible. In this context, our paper bridges the disciplinary gap between agricultural economics and communication studies by demonstrating how to validate and identify search terms for content analysis in agricultural economics applications. Doing so, we provide a brief introduction to the conceptual background (“performance metrics”) of search term validation (Section 2). Based on this, we describe the step-by-step approach to search term validation in ACA and, for comparison purposes, two ad hoc approaches based on researchers’ intuition and artificial intelligence (AI) using the topic “animal welfare” as an example (Section 3). After the comparison of the performance of all search terms (Section 4), we discuss the limitations of our findings and their implications for future ACA applications (Section 5).

2. Conceptual background and performance metrics

Systematic search term validation progressively refines an initial search term, search phrase, or search string. This stepwise approach is based on the “systematic evaluation of search phrases, with estimates of their ability to return relevant stories and reject extraneous ones” (Stryker et al. 2006: 414). The core idea is to use sample-based estimates for precision and recall as metrics for the (hopefully increasing) validity of search terms as they are progressively refined when moving from a very “open search term” to a more “narrow search term” and finally a ‘“closed search term”. “Precision” estimates the proportion of retrieved text data that is relevant when a particular search term is used to extract data from the text

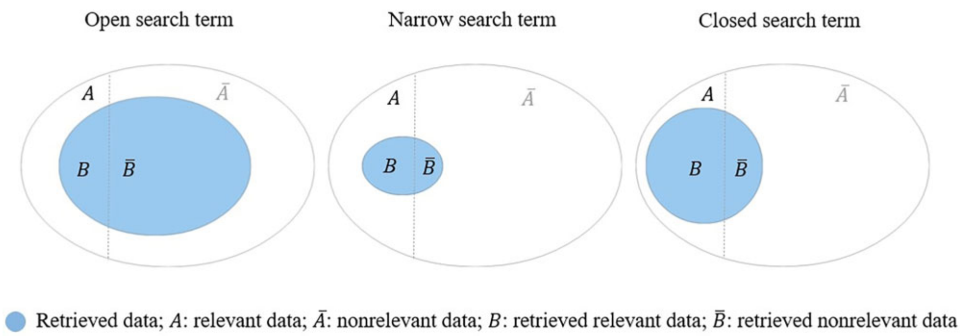

Figure 1. The concepts of precision and recall.
Source: Authors' representation based on Stryker et al. (2006).

universe. "Recall" estimates the proportion of relevant text data that is retrieved when that search term is used (cf. Lacy et al. 2015, Riffe et al. 2019). Figure 1 visualizes the underlying concept:

The text universe (e.g. the digitized archives of the selected newspapers over the defined period of time) represents the entirety of the data entering the analysis. This universe comprises not only relevant data (A) but also non-relevant data ($\bar{A}$). The subset that is retrieved from the universe by a certain search term contains only a subset of the relevant data (B) and misses a part of the relevant data. It also contains non-relevant data or noise ($\bar{B}$). Referring to Fig. 1, precision p (i.e. the share of retrieved texts that are relevant) and recall r (i.e. the share of relevant texts that are retrieved) can be defined as follows:

$$p = \frac{B}{B + \bar{B}}, \tag{1}$$

$$r = \frac{B}{B + A}. \tag{2}$$

The harmonic mean of both metrics—called "$F1$ score"—is used as a single summary performance index of the validity of a search term (Stryker et al. 2006, Lacy et al. 2015, Mahl et al. 2022).

$$F1 = \frac{2}{\frac{1}{p} + \frac{1}{r}} = \frac{2pr}{p + r}. \tag{3}$$

Using the harmonic mean implies that increases in precision and recall cause decreasing increases of the single performance index $F1$. For example, $p = 0.8$ and $r = 0.8$ leads to a higher $F1$ score (0.8) compared to, say, $p = 0.7$ and $r = 0.9$, which leads to $F1 = 0.7875$.

3. Methods of search term generation

3.1 Systematic search term validation: an overview of working steps

Using the topic "animal welfare", we demonstrate in this section how the systematic validation of search terms can be carried out to provide a high-quality and transparent database for ACA. The description is guided by the following questions: (1) How can we obtain reliable relevance criteria that effectively differentiate between relevant and non-relevant data? (2) How can we obtain valid search terms that retrieve articles with high precision and recall?

3.1.1 Specification of the context and the text universe

After specifying the public discourse and research question, the first step in content analysis is to define and justify the text universe that will be used for the analysis (cf. Mahl et al. 2022). To be more precise, a decision must be made as to which publication forms (e.g. journal, newspaper, social media, etc.), which publication outlets, and which publication periods are to be included in the text universe.

In our exemplary case, we looked at German newspaper articles published between 2010 and 2023. The study investigates how German media frame the association between intensive livestock farming and violations of the German animal protection act. The analysis follows Entman's (1993: 52) definition of frames according to which communication texts highlight some aspects of a perceived reality and thus "promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation" (Entman 1993). The promoted aspects of reality are called frame elements.[1]

Not all available newspapers were suitable to provide information about the public discourse at a national level—for example, due to strong regional limitations or very low circulation. We therefore first selected suitable newspapers to be included in the analysis. As national newspapers can be considered opinion leaders among all media (Jarren and Donges 2011), we focused on the German national newspapers with the highest circulation as well as the most widely read German agricultural trade magazine. For our analysis, we selected the weekly papers *DIE ZEIT*, *Der Spiegel*, *Stern*, and *FAZ* as well as the daily papers *tageszeitung (taz)* and *Die Welt*. We also included *Die BILD* (including *BILD am Sonntag*), the most popular tabloid in Germany, and *agrarzeitung*, the agricultural trade magazine with the highest circulation. This text universe was chosen to cover a broad spectrum of opinions and a wide range of social groups. Politically, *taz* and *DIE ZEIT* can be considered left-liberal, while *FAZ*, *Die Welt* (market-liberal), *Stern* (liberal), and *Der Spiegel* are more conservative. The national daily and tabloid newspaper with by far the highest circulation is the right-leaning *BILD* (IVW 2022). We also included *agrarzeitung*, as it is one of Germany's leading outlets for agricultural news from Germany, Europe, and around the world. We included all articles from 2010 to 2023. All articles, except those of the FAZ, can be retrieved from the online platform "LexisNexis" via search terms. The FAZ has its own online archive.

3.1.2 Selection of potentially relevant texts using an open search term

The first step in search term validation is to create an open search term that identifies a large portion of the relevant articles and promises to produce a "condensed" working dataset in which the proportion of relevant text passages is substantially higher than in the text universe or a simple random sample from the text universe. Commonly, it begins with a search string containing known synonyms, metaphors, and alternative combinations of the words used in the research question. The open search term is then enriched by adding terms found in an informal review of the texts already retrieved (Lacy et al. 2015). Boolean operators are generally used to link different expressions. Using the AND-operator means that the linked words or phrases must both occur, regardless of how close or far apart they are or in which order they appear in the text. The restriction to simultaneous occurrences limits the amount of data retrieved but also allows search results to be focused on a specific topic—that is, it increases precision. Using the OR-operator, in contrast, means that data is retrieved whenever one of the linked synonyms, alternative forms of expression, acronyms, etc. occurs anywhere in the text. Using the OR-operator thus increases the amount of retrieved data and increases recall (Stryker et al. 2006). In addition, using "wild cards" that take the place of a single character or of several characters provides flexibility regarding the variability of verbal expressions that are to be retrieved.

Table 1. Inductively and deductively generated elements of the open search term.

Inductively generated elements	Deductively generated elements		
	Synonyms	Literature	First review of retrieved articles
tiersch*tz! (relating to different versions of "animal protection" or "animal protector")	–	–	–
AND		OR	
landw! (relating to different versions of "agriculture" and "farmer")	agrar!, agro!, b*uer!, agrarier!, agronom! (relating to different synonyms of "agriculture" and "farmer"	schweinehalt!, schweinem*st!, geflügelhalt!, milchviehalt! (relating to different versions of "pig farming", "poultry farming", and "dairy farming')	Legehenn!, rinderhalt!, rinderm*st! (relating to different versions of "laying hen farming" and "cattle farming")

Table 1 shows the composition of the open search term used in our content analysis on animal welfare. It consisted of "inductive" terms used in the research question (1st column), "deductive" terms (i.e. their synonyms) (2nd column), terms used in previous studies by other authors (3rd column), and terms detected by our first informal review of articles (4th column).

For the texts that could be accessed through LexisNexis, our search term was specified using the Boolean operators AND and OR as well as the wildcard symbol asterisk (*), which replaces a single character anywhere in a word except for the first character. Additionally, we used the truncation symbol exclamation mark (!), which replaces any number of characters at the end of a word and allowed us to search for alternative word endings. For example, the term tiersch*tz! could match the German words Tierschutz, Tierschützer, or Tierschützerin, relating to different versions of "animal protection" and "animal protector". We also used the question mark (?), which replaces the number of characters equal to the number of question marks.[2]

Combining the search term elements listed in Table 1, the open search term that we used for LexisNexis was as follows:

> tiersch*tz AND (landw! OR agra! OR agrarier! OR agro! OR agronom! OR b*uer! OR geflügelhalt! OR geflügelm*st! OR legehenne! OR milchviehhalt! OR tierhalt! OR rinderhalt! OR rinderm*st! OR schweinehalt! OR schweinem*st!).[3]

3.1.3 Specification of reliable relevance criteria

Before the performance of a search term can be validated using precision and recall, the coding criteria ('relevance criteria') used to differentiate between relevant and non-relevant texts must be defined. Figure 1 illustrates the task. Without reliable criteria, it would not be possible to classify a retrieved text as relevant (an element of B) or nonrelevant (an element of $\bar{B}$). As those criteria must be defined by the individual researcher, it is crucial to find out how reliably they differentiate between relevant and non-relevant texts.

The development of reliable relevance criteria is also a multistep procedure. "Reliable" means that different human coders using these criteria independently arrive at a minimum number of consistent decisions for the same texts (Lacy et al. 2015). As a general rule of thumb, the texts used to determine "intercoder reliability" should be read independently by a minimum of two coders (Kolbe and Burnett 1991, Krippendorff 2004a, Stryker et al. 2006, O'Connor and Joffe 2020). Intercoder reliability needs to be quantified. An established measure is Krippendorff's α (Krippendorff 2004a, Hayes and Krippendorff 2007). It ranges from 0 (absence of reliability) to 1 (perfect reliability) and can be used for any number of coders (Hayes and Krippendorff 2007). Commonly, $\alpha = 0.8$ is used as stopping rule (i.e. the reliability threshold that relevance criteria must exceed to be accepted as reliable) (Lombard et al. 2002, Krippendorff 2004b, Neuendorf 2017).

The systematic identification of relevance criteria begins with a draft of criteria that are derived inductively—similar to the formation of the open search term—from the research question or the objectives or from criteria that already exist in the literature. In addition, further criteria can be created deductively from an initial informal review of some of the texts obtained through the open search term. To check how reliable these criteria are, at least two people independently code a random sample of texts found using the open search term into relevant and nonrelevant. If the threshold for Krippendorff's α is not met, the existing criteria must be refined. Necessary modifications or the incorporation of new criteria can be discussed with the coders, who can provide insights based on their experiences from the coding process.

In our content analysis on animal welfare, we initially formed six criteria inductively from the research question. One example of an initial inductive relevance criterion was: "The text is relevant if at least one contextual association is established between the terms 'tiersch*tz' and 'landw!' or one of its synonyms." Another criterion stated that texts are not relevant if the expressions of the search terms are "mentioned in passing, comparatively or remuneratively". The initial relevance criteria were then reviewed by human coders who read texts obtained using the open search term. Due to the large number of texts that we found using the open search term (2,699), we followed the procedure of Stryker et al. (2006) and drew random samples that could actually be read by the human coders with reasonable effort.

As two or more coders are generally recommended, we recruited[4] three coders for each iteration, who, in each iteration, independently coded a sample of 100 newspaper articles. Each iteration of the review process was carried out with a new random sample of $n = 100$ from the 2,699 texts that had been initially retrieved by the open search term. All coders were recruited from the social environment of the researcher without requiring expertise in coding, content analysis, or animal welfare. The only requirement for recruitment was a university degree and attendance of the coder training in which the task and the relevance criteria were introduced in detail. The coding results were received anonymously and processed for evaluation purposes.

At the end of each iteration, we used R to determine α. We applied the decision rule $\alpha \geq 0.8$ to stop further iterations. To be more precise, as long as intercoder reliability was not satisfactory ($\alpha < 0.8$), we carried out further iterations to generate more reliable relevance criteria by eliminating, adjusting, or adding further criteria. After each iteration, the coders could suggest improvements to the relevance criteria. The final version of relevance criteria was reached when intercoder reliability was satisfactory and $\alpha \geq 0.8$.

3.1.4 Specification of a final search term with satisfactory performance

The final step in the systematic validation of search terms is to evaluate their performance based on their precision p and their recall r (in addition, $F1$ is commonly displayed to provide readers with an accessible summarized score for the overall validity of a search term). Similar to the stopping rule in the search for reliable relevance criteria (cf. Section 3.1.3), a threshold is useful to determine when to stop searching for a more valid search term (Mahl et al. 2022). As with other thresholds used in content analysis, this is not a question of right or wrong. Instead, researchers must make and justify stopping decisions depending on the research context. Aspects to consider are, for example, a concept's latency and the ambiguity of related terms. For concepts that are more latent and difficult to grasp with search strings (e.g. sentiment), lower recall and precision values may be acceptable (van Atteveldt et al. 2021). For concepts where a finite list of words is easier to identify (e.g. country mentions), higher recall and precision values can be targeted. If a concept is highly ambiguous or has multiple meanings (e.g. "bank," which can refer to a financial institution or the side of a river), a lower threshold for recall might be acceptable to ensure that all relevant instances are captured, even at the cost of including some irrelevant results.

Once the thresholds have been defined by the researcher, search terms can be continuously refined—starting with the open search term (cf. Section 3.1.2) and using the relevance criteria specified earlier (cf. Section 3.1.3)—until texts are retrieved with an acceptable level of precision and recall. This estimation is commonly based on the last sample that was used in the search for the final relevance criteria. For this sample, the "true" ratio of relevant and non-relevant articles is known because the human coders actually read all texts. Consequently, precision and recall and thus the quality of the categorization of texts into relevant and nonrelevant that is generated by the search term can be determined.

Refining the search term means improving its performance by modifying its content and structure. On the one hand, this can be achieved by adding presumably appropriate terms as well as by removing or excluding inappropriate ones. On the other hand, the Boolean operators and wildcards (cf. Section 3.1.2) can be adjusted to systematically increase recall, precision or both. New terms can be taken from the coders' feedback in the previous validation step. After each refinement, the performance is measured again. This procedure has to be repeated until the predefined stopping rule is reached.

In our content analysis, we used the label "Sample I" to denote the last random sample ($n = 100$) from the 2,699 texts retrieved using the open search term that we used in the search for the relevance criteria. We used the label "search term candidate" for subsequent versions of the search term that were tested and successively refined based on their performance in Sample I. We could monitor the performance of each search term, as we were able to estimate precision by counting in Sample I how many relevant articles (elements of B) and how many non-relevant articles (elements of $\bar{B}$) were retrieved. We were also able to estimate recall because we knew how many relevant articles (elements of A) were not retrieved (cf. Figure 1). We set the threshold for both precision and recall at 0.8, a value also used in the Mahl et al. (2022) study on climate change. We chose these thresholds as we found that the concept of "animal welfare" is comparably difficult to capture with search terms as the concept of "climate change." As soon as the predefined performance thresholds of 0.8 for both precision and recall were achieved, we stopped the search term refinement process.

In our case, refinement was achieved through ad hoc adjustments of successive search term candidates that had not yet led to a satisfactory performance. For example, we added terms that had been found as side products by the coders during the earlier search for reliable relevance criteria. We also removed terms if they had a negative effect on performance. In some cases, we adjusted the wildcards to better control the variations of terms that were included in the search term. After each adjustment, the new search term candidate was used in the text universe and then tested by determining which of the known relevant and non-relevant articles in Sample I it had been able to retrieve. Quantifying precision and recall for each search term candidate showed which adjustment steps led to which performance effects. It also showed when the quest for a better search term could finally be stopped according to the stopping rule.

The final search term whose performance in Sample I was deemed satisfactory according to the stopping rule was finally subjected to an additional performance test, as we could not exclude the possibility that it only worked well in the initial Sample I of size $n = 100$.

Therefore, we drew a new random sample ("Sample II") of size $n = 200$ from the 2,699 texts that had been retrieved using the open search term. The additional test basically repeated the above-mentioned steps of applying the search term to the text universe and then monitoring its performance in a sample. Since we used a new sample, the coders had to read these new articles and separate them into relevant and nonrelevant based on the predefined relevance criteria. Finally, if precision and recall of the final search term also reached at least 0.8 in Sample II, we considered this to be sufficient confirmation of the validity of the search term.

3.2 Unsystematic search terms generation: a demonstration example

As systematic search term validation is still not commonly used in content-analytic studies, it is interesting to learn how non-systematic search term generation based on intuitive or AI-assisted approaches perform compared to a rigorous systematic approach. Non-systematic approaches have cost advantages as they require relatively little research resources including time. The crucial question is therefore how well they discriminate between relevant and irrelevant data. To address this question, we compared the precision and the recall of an intuitive and of two AI-assisted approaches with systematic search term validation.

The intuitive search term was formed in just one step from the researcher's ad hoc associations regarding the topic and ideas regarding the combination of terms with which as many of the potentially relevant articles could presumably be found. This procedure is essentially comparable to other ad hoc searches such as those involved when using search engines like Google. We have decided to link the terms "tiersch*tz!" and "landw!" using the operator AND. The underlying rationale was to retrieve all articles from online data bases which deal to some extent with protection of animals ('"tiersch*") and agriculture or farmers ("landw*"). We denote this search term as unsystematic because there is no check of reliability (e.g. would another person choose the same terms?) and validity (share of relevant and non-relevant articles).

To create the AI-supported search terms, two researchers were asked to independently but simultaneously prompt ChatGPT (4.0) and Gemini (1.5 Flash) to obtain a comprehensive search term from each chatbot. The only restrictions were a time limit of 15 min per prompting and the requirement that both researchers should use their individual starting-prompt for both large language models (LLM's). The researchers were given general advice regarding effective prompt development (which we received from ChatGPT and Gemini themselves). Both researchers started with ChatGPT and continued with Gemini. As each of the two researchers suggested a search term per LLM, we determined a consensus search term per chatbot in a group discussion between the prompters and a moderator. Thus, we ended up with two AI-supported search terms—one consensus search term from ChatGPT and one consensus search term from Gemini. As they were not further tested for reliability and validity, they are classified as being derived from an unsystematic approach.

Finally, we compared the performance metrics of the validated, intuitive search terms and the two AI-assisted search terms. Since the performance of the validated search term had already been determined, we applied the unsystematic search terms to Sample I in a manner analogous to the monitoring of search term refinements in Section 3.1.4. In this way, we were able to determine the extent to which the unsystematic search terms based on Sample I find the relevant articles and reject the irrelevant ones.

4. Performance of systematic search term validation compared with approaches based on researchers' intuition and AI

In the following section, we present the results of the systematic approach described above. This includes a detailed account of the final relevance criteria, a comprehensive account of the step-by-step development of the final valid search term, and a comprehensive review

Table 2. The identification of reliable relevance criteria within three iterations.

	Random samples ($n = 100$) drawn (n) from the population of 2,699 articles found using the open search term in 10 newspapers	Articles coded as relevant (C1,…, C9 = Intercoder)			Krippendorff's α
Iteration 1	$n_1 = 100$ (10 articles per newspaper)	C1: 42	C2: 52	C3: 43	0.65
Iteration 2	$n_2 = 100$ (10 articles per newspaper)	C4: 43	C5: 48	C6: 64	0.61
Iteration 3	$n_3 = 100$ (10 articles per newspaper)	C7: 62	C8: 63	C9: 61	0.80

of the performance metrics of each search term candidate. Furthermore, we address the question of whether the systematic search term validation procedure indeed provides more valid search terms than approaches that do not comply with this procedural standard. We contrast the performance metrics (precision p, recall r, and the overall measure $F1$) obtained via systematic search term validation with those of an "intuitive approach" and an "AI-assisted approach."

4.1 Relevance criteria

As mentioned in Section 3.1.2, reliable relevance criteria are required before the search term validation can start. In an iterative process, we refined the initial list of criteria and monitored their consistency when used by different human coders independently via the intercoder reliability. With each iteration, there were decisive changes to the relevance criteria. For example, the number of criteria increased from six in the first iteration to 13 in the third iteration (see the list of criteria in Appendix B). Some criteria were eliminated because the coders indicated that these criteria were either not helpful or only marginally helpful for the categorization of texts into relevant and nonrelevant. Conversely, other criteria were added or existing criteria refined. All modifications were derived from the feedback survey, which was completed by each coder at the conclusion of the coding process. After the third iteration, we judged the relevance criteria to be reliable, as we obtained an intercoder reliability (Krippendorff's α) of $\alpha = 0.8$, which we used as a stopping rule (cf. Table 2).

Since we needed three iterations to reach the threshold $\alpha = 0.8$, a total of nine human coders read 300 newspaper articles to form the relevance criteria. The size of the newspaper articles ranged from a quarter page to four pages of pure text (in .txt format). The survey of the processing time revealed that it took an average of 3 min per text to read and make the final coding decision (i.e. 5 h to complete the entire coding task).

4.2 Validated search terms

This section describes the results of the step-by-step procedure of validating the search terms in our exemplary content analysis on animal welfare. This validation procedure was based on Sample I, which was the label that we attached to the last random sample ($n_3 = 100$) from the initially retrieved 2,699 texts that was used in the earlier search for reliable relevance criteria. According to the results of this earlier step, we considered 61 articles in this sample to be relevant, as all three coders had agreed in their relevance judgement of these articles (cf. Table 2).

Validation means monitoring the performance metrics of search term candidates as they are successively refined. Table 3 summarizes the monitoring results and shows how

Table 3. Successive evaluation of search term candidates using sample I (n = 100, with a share of 61 relevant and 39 non-relevant texts).

Search term versions	No. of retrieved texts	No. of retrieved relevant texts in Sample I	No. of retrieved non-relevant texts in Sample I	p[(a)]	r[(a)]	$F1$[(a)]
Open search term (A)	2,699	–	–	–	–	–
Search term candidate B	3,070	55	32	0.632	0.902	0.743
Search term candidate C	3,841	41	12	0.774 (+0.142)	0.672 (−0.230)	0.719 (−0.024)
Search term candidate D	4,551	53	15	0.779 (+0.006)	0.869 (+0.197)	0.852 (+0.102)
Final search term (E)	**4,399**	**52**	**11**	**0.825 (+0.046)**	**0.852 (−0.016)**	**0.839 (+0.017)**

[a]Values in brackets indicate the incremental change in the performance metrics achieved in the step-by-step refinement of term candidates compared to the previous candidate.

precision p, recall r, and the $F1$-score gradually changed with the move from one search term candidate to the next. Starting with the open search term, which had been designed to produce a "condensed" working dataset that contains a large share of the relevant texts, four gradual refinements (B, C, D, and E) were necessary to bring about the final search term E. With $p = 0.825$, $r = 0.852$, and $F1 = 0.839$, search term E showed satisfactory performance according to the predefined performance threshold (stopping rule), according to which estimated precision and recall should be at least 0.8.

The individual refinement steps (for all refinement details look at Appendix D) that underlie the gradual performance improvement of the search term candidates (see details of the search terms in Appendix C) were made by adding and excluding terms as well as by exchanging truncation symbols. To provide a few illustrative examples: (1) moving from candidate search term B to C, we used the BUT NOT-operator to exclude non-relevant texts such as those dealing with zoo or pets. (2) The performance increase from candidate search term C to D was obtained by adding terms for breeding as well as other animal species and production areas. In addition, "insects!" was removed from the BUT NOT-operator as many relevant articles that mentioned insects—for example, when discussing alternatives to animal food production—had been erroneously excluded by search term C. (3) The decisive increase in performance toward the final search term E was achieved through various further refinements, including those that counteract semantic ambiguity through truncation symbols. For example, "b*uer!" also identified texts dealing with professional fire departments (in German: "Berufsfeuerwehr"). To ensure that texts dealing with fire departments are not erroneously included, we replaced the wildcard symbol (*) in "b*uer!" by the question mark (?), thus reducing the number of characters that can be replaced to one.

As mentioned in Section 3.4, we carried out an additional test of the final search term by determining its performance in another random sample (Sample II; $n = 200$) drawn from the 2,699 texts initially identified by the open search term. According to the categorization of texts in Sample II done by our human coders, 127 out of the 200 articles were relevant, and 73 articles were not. Search term E, which retrieved 116 relevant articles and 28 non-relevant articles, thus showed precision $p = 0.806$, recall $r = 0.913$, and $F1 = 0.856$. The performance estimated from Sample II is essentially the same as that estimated from Sample I. We considered this consistency to be sufficient confirmation of the validity of search term E.

Adherence to the standards of systematic search term validation allows researchers to transparently demonstrate how a search term used in a content analysis was determined and what discriminatory performance it achieved. While this is a value in its own right, the

Table 4. The performance of search term E, an intuitive search term, and two AI-assisted search terms based on the evaluation in Sample I ($n = 100$, with a share of 61 relevant and 39 non-relevant texts).

	No. of retrieved texts	No. of retrieved relevant texts in Sample I	No. of retrieved non-relevant texts in Sample I	p	r	$F1$
Final search term E	4,399	52	11	0.825	0.852	0.839
Intuitive search term	1,009	47	18	0.723	0.770	0.746
ChatGPT (4.0)	2,228	29	6	0.829	0.475	0.604
Gemini	733	10	4	0.714	0.164	0.267

question remains as to whether systematically derived search terms also have more discriminatory power than intuitively ad hoc formed search terms. Given the rapid recent progress and growth of AI, in Table 4, we looked into this issue by comparing the systematically identified search term E with an intuitively formed search term and two AI-assisted search terms (see Appendix C for search term details).

Neither the intuitive nor the AI-assisted search terms reached the threshold of 0.8 for the metrics r and p at the same time. While the performance of the intuitive search term ($p = 0.723$, $r = 0.770$, and $F1 = 0.746$) was below that of the final search term E, it was still close. In fact the AI-assisted search term by ChatGPT exceeded the threshold regarding precision with $p = 0.829$, but $r = 0.475$, and $F1 = 0.604$ were still unsatisfying. In contrast, the performance of the AI-assisted search term by Gemini was much worse, with performance metrics as low as $= 0.714$, $r = 0.164$, and $F1 = 0.267$.

Table 4 illustrates that a one-sided view of the absolute number of texts found would be misleading when assessing the performance of a search term. If we were to follow this criterion, we would erroneously favor the AI-supported search term by ChatGPT (2,228 texts) over the intuitive search term (1,009 texts). This difference in the number of retrieved texts arises because the AI-assisted search term is much more complex. However, the proportion of relevant articles (29) and thus the discriminatory power is furthest away from the threshold value. Conversely, the performance of the intuitive search term is close to that of the final search term. With regard to the number of articles retrieved from Sample I, it could be argued that the discriminatory power of the intuitive search term is acceptable. However, if we consider what a precision of $p = 0.723$ means for large datasets in the ACA, we can estimate that for a universe of, say, 10,000 texts, applying the intuitive search term could cause 2,770 non-relevant texts to slip into the analysis unobserved as noise.

5. Discussion

With the technological advances that have greatly increased the availability of large digitized text datasets as well as the ability of researchers to analyze them, ACA has found its way into many disciplines, including agricultural economics. Despite the presence of computer-assisted ("automatic") data processing, the challenge of specifying search terms that effectively differentiate between relevant and non-relevant texts remains. However, the growing popularity of content analysis does not always mean that the standards for transparent search term validation are adhered to. As ACA is likely to be increasingly used in the social sciences, bridging disciplinary communication gaps and disseminating good practice standards is of great importance. This is the starting point of this paper, which used an exemplary content analysis on the topic of animal welfare to illustrate the procedural steps of systematic search term validation that are necessary to obtain a transparent, high-quality database.

The systematic search term validation provided some interesting findings. First, the discriminatory power of the candidate search terms as they were successively refined could be gradually increased until they met the stopping rule defined in this study, with both estimated precision $p = 0.825$ (i.e. the proportion of retrieved text data that is relevant) and estimated recall $r = 0.852$ (i.e. the proportion of relevant text data that is retrieved). The resulting overall discriminatory power measured via the harmonic mean ($F1$-score) was 0.839. Second, the discriminatory power of this systematically identified search term outperformed an intuitively formed search term, with an overall discriminatory power of only $F1 = 0.746$. The difference was even more pronounced when compared to a second and third non-validated search term specified using AI. Here, the overall discriminatory power was only $F1 = 0.604$ by ChatGPT and $F1 = 0.267$ by Gemini. Third, and almost more importantly, the decisive advantage of the systematic validation of search terms is that it is transparent, in contrast to the use of non-validated search terms. Even if a subsequent review analysis were to find that a non-validated search term had satisfactory discriminatory power, the discriminatory power or relevance criteria (and their reliability) used to assess the validity of the search term would remain unclear. Consequently, it would be impossible to either fully understand and retrace the results or replicate the study, which is increasingly considered to be unacceptable in modern research.

Ad hoc adjustment during search term refinement itself could be considered as prone to subjectivity. For instance, subjective term picking could have affected the efficiency of the adjustments, for example, the actual number of iterations needed by the researcher to meet the desired threshold. This particularly matters at the beginning of the refinement process with low performance metrics, as the researcher is "swimming in the ocean of terms" with limited insight and orientation on which conceptual changes will affect performance. However, as long as two hypothetical individuals independently arrive at the same search term for an identical topic despite taking different paths, subjectivity is unlikely to affect the performance metrics (which are objective measures) and search terms themselves. It is also possible for ad hoc terms to be included in the search term because they were inspired by reading relevant texts. This is not inherently problematic. However, issues arise when terms are included solely because they frequently occur alongside relevant search terms in the texts, without any connection to the research question. Even if performance improves during the refinement process in the sample, the final search term may become misleading in the broader text universe.

While no specific refinement approaches can be found in the literature, we aimed to reduce subjective biases by systematically asking all coders to suggest changes to the search term. Thus, after each refinement and coding process, we received independent input from three coders on advisable changes based on their experience after reading 100 articles each. As we received their suggestions via an open survey, we made the refinements based on a qualitative assessment. We see potential for future studies to strengthen the reliability of such refinements. For example, coders could be asked for specific and complete search term variants (rather than just partial changes) and the researcher could then test how consistent the different suggestions are. In addition, systematic group discussions between the coders and the researcher could be used to reach a more objective cross-personal consensus on the search term changes.

6. Conclusion and future research

The relevance of search term validation is clearly supported by the current findings. To the best of our knowledge, our study is the first systematic attempt to bridge the disciplinary gap between agricultural economics and communication studies by demonstrating how to validate and identify search terms that are able to generate valid databases for content analysis in agricultural economics applications. The procedural standards described in this paper help users of content analysis, such as agricultural economists, mitigate issues with transparency and the replicability of content analytic studies. Disclosing which terms are included in a search term and which are not facilitates the communication of research results and their intelligibility. Moreover, content analysis, such as systematic literature reviews, is an established instrument to both generate research hypotheses and inform policy-makers. Our paper provides additional tools to strengthen these outputs.

To further assess the generalizability of the proposed approach, future research should address other regions (including different languages) and contexts within the agricultural domain. It would also be interesting to explore different datasets, such as social media content. Additionally, future research should analyze the role of coder expertise. Regularly examining the potential of AI would also be valuable, as considerable changes can be expected in the future. The results of content-analytic studies may depend on the open search term that was originally used to obtain the "condensed" working dataset. It could therefore be useful to investigate whether degrees of freedom (flexibility) in specifying the open search term jeopardize the robustness and quality of results.

In view of the rapid progress and growth of AI, a host of interesting questions finally arise regarding the competitiveness of AI in terms of its discriminatory power compared to the structured process of systematic search term validation: (1) Are there other prompts that would give better results when using a particular AI chatbot? (2) How would multiple AI chatbots perform in comparison at a given point in time? (3) As AI chatbots evolve over time, would different (better) results be achieved and could they potentially outperform systematic search term validation in the future? (4) How could the transparency necessary for progress in research be ensured in AI applications? Numerous systematic tests of AI chatbots in content-analytic studies would be needed to answer these questions and thus contribute to assessing the quality of results obtained by a black box approach such as AI.

Acknowledgments

We acknowledge the financial support of the Open Access Publication Fund of the Martin Luther University Halle-Wittenberg. The authors declare that they have no relevant or material financial interests related to the research described in this paper. The study has been approved by the institutional review board (IRB) of Martin Luther University Halle-Wittenberg. We thank Tony Beyer, Sam Müller, Gil Noack, Alexandra Ratke, Konrad Bretschneider, Ronja Noethen, Kai Rogge, Anne-Marie Heger, Karoline Vorlop, Alexander Groß, Lucie Sophie Schirmer, and Sandra Tappendorf for their valuable contributions to coding tasks and discussions. We would also like to thank Michael Reiss (HBI) as well as Jakob Moritz-Eberl, Svenja Schäfer, and Teresa Weikmann from the Institute of Communication at the University of Vienna for their valuable and insightful contributions from a communication studies perspective. Furthermore, we would like to extend our appreciation to the PhD Club at the University of Vienna and its participants, who played a pivotal role in fostering interdisciplinary dialogue between agricultural economics and communication studies.

Conflict of interest

None declared.

Data availability

The text data is freely available at the online portals LexisNexis (https://advance-1lexis-1com-1zf6r3soj03de.erf.sbb.spk-berlin.de/bisacademicresearchhome?crid=b5965322-6c2d-4b23-ac00-cbf63576c7ff&pdmfid=1516831&pdisurlapi=true) and F.A.Z-Bibliotheksportal (https://www.faz-biblionet.de/faz-portal).

End Notes

1 The analysis was based on texts at the article level. We used R to analyze the newspaper articles automatically. This enabled us to include a large amount of data in the frame analysis.
2 Different symbols are used for wildcards in the FAZ Library Portal. There, the question mark (?) replaces the asterisk (*), and the asterisk (*) replaces the exclamation mark (!). Apart from that, terms and operators are identical.
3 Because the wildcard symbols are different in the FAZ Library Portal, the open search term for the FAZ texts was as follows: tiersch?tz AND (landw* OR agra* OR agrarier* OR agro* OR agronom* OR b?uer* OR geflügelhalt* OR geflügelm?st OR legehenne* OR milchviehhalt* OR tierhalt* OR rinderhalt* OR rinderm?st OR schweinehalt* OR schweinem?st*).
4 The work with human subjects has been approved by the institutional review board of Martin Luther University Halle-Wittenberg.

Appendix A. Dictionary of terms for search term E: German—English

Terms are sorted alphabetically, not in the order in which they appear.

German term	Finds (maximum 3 examples each)	*Translation*
aal!	Aal Aalräucherei Aalen (Ort)	*Eel* *Eel smokehouse* *Aalen (city name)*
agra!	Agrarindustrie Agrarbranche Agrarpolitik	*Agricultural industry* *Agricultural sector* *Agricultural policy*
agro!	Agrosystem Agronomin Agronomisch	*Agro-system* *Agronomist (female)* *Agronomic*
aquakultur!	Aquaklulturproduktion Aquaklulturbetrieb Aquakulturprodukt	*Aquaculture production* *Aquaculture operation* *Aquaculture product*
aquari!	Aquarium Aquarianer Aquarien	*Aquarium* *Aquarist* *Aquariums*
b*uer!,	Bauer Bäuerin bäuerlich	*Farmer* *Farmer (female)* *Agricultural*
ferkel!	Ferkelerzeuger Ferkelzucht Ferkelmast	*Piglet producer* *Piglet breeding* *Piglet fattening*
fischw!	Fischwirtschaft Fischwirt Fischwilderei	*Fisheries industry* *Fish farmer* *Fish poaching*
fischz*cht!	Fischzucht Fischzüchter Fischzuchtanlage	*Fish farming* *Fish breeder* *Fish farming facility*
geflügelhalt!	Geflügelhaltung Geflügelhalterin Geflügelhalter	*Poultry farming* *Poultry farmer (female)* *Poultry farmer*
geflügelm*st!	Geflügelmastanlage Geflügelmastbetrieb geflügelmäster	*Poultry fattening facility* *Poultry fattening operation* *Poultry fattener*

German term	Finds (maximum 3 examples each)	*Translation*
geflügelprod!	Geflügelprodukte Geflügelproduktion Geflügelproduzent	*Poultry products* *Poultry production* *Poultry producer*
geflügelz*cht!	Geflügelzucht Geflügelzüchter Geflügelzucht-Verein	*Poultry breeding* *Poultry breeder* *Poultry breeding association*
h?hn!	Hahn Huhn Hühner	*Rooster* *Chicken* *Chickens*
hähnchenhalt!	Hähnchenhaltung Hähnchenhalter EU-Hähnchenhaltungsrichtlinie	*Chicken farming* *Chicken farmer* *EU Chicken farming directive*
hähnchenm*st!	Hähnchenmast Hähnchenmastanlagen Hähnchenmastbetrieb	*Chicken fattening* *Chicken fattening facilities* *Chicken fattening operation*
hähnchenz*cht!	Hähnchenzucht Hähnchenzüchtung Hähnchenzüchter	*Chicken breeding* *Chicken breeding* *Chicken breeder*
hamster!	Hamster Hamsterkäfig Hamsterrad	*Hamster* *Hamster cage* *Hamster wheel*
haustier!	Haustierhaltung Haustier Haustiere	*Pet keeping* *Pet* *Pets*
hund?	Hundehalter Hundehalterinnen Hund	*Dog owner* *Dog owners (female)* *Dog*
k*lb!	Kälberaufzucht Kalb Kälbermast	*Calf rearing* *Calf* *Calf fattening*
katze!	Katzenhaltung Katze Katzenhalter	*Cat keeping* *Cat* *Cat owner*
kleintier!	Kleintierhaltung Kleintierhalter Kleintierhalterinnen	*Small animal keeping* *Small animal keeper* *Small animal keepers (female)*
küken!	Küken Kükentöten Kükentransport	*Chicks* *Chick culling* *Chick transport*
landw!	Landwirt Landwirtschaft Landwirtschaftlich	*Farmer* *Agriculture* *Agricultural*
Legehenn!	Legehennenproduktion Legehenenhaltung Legehennenhalter	*Laying hen production* *Laying hen keeping* *Laying hen keeper*
leser*brief!)	Leserbrief Leserinnenbrief Leserbriefe	*Letter to the editor* *Letter to the editor (female)* *Letters to the editor*

German term	Finds (maximum 3 examples each)	*Translation*
masth*hn!	Masthuhnhaltung Masthühner Masthuhninitiative	*Broiler farming* *Broiler chickens* *Broiler chicken initiative*
mastschwein!	Mastschweine Mastschweinebestand Mastschweinestall	*Fattening pigs* *Fattening pig stock* *Fattening pig barn*
milchviehalt!	Milchviehaltung Milchviehhalter Milchviehhalterinnen	*Dairy farming* *Dairy farmer* *Dairy farmers (female)*
milchviehprod!	Milchviehproduzent Milchviehproduktion Milchviehproduzentin	*Dairy producers* *Dairy production* *Dairy producer (female)*
milchviehz*cht!	Milchviehzucht Milchviehzüchter Milchviezuchtbetrieb	*Dairy breeding* *Dairy breeder* *Dairy breeding operation*
nutztierhalt!	Nutztierhaltung Tierschutz-Nutztierhaltungsverordnung Nutztierhalter	*Livestock farming* *Order on the protection of animals and the keeping of production animals* *Livestock farmer*
rinderhalt!	Rinderhaltung Rinderhalter Rinderhalterinnen	*Cattle farming* *Cattle farmer* *Cattle farmers (female)*
rinderm*st!	Rindermast Rindermäster Rindermastbetrieb	*Cattle fattening* *Cattle fattener* *Cattle fattening operation*
rinderprod!	Rinderprodukte Rinderproduzenten Rinderproduktion	*Cattle products* *Cattle producers* *Cattle production*
rinderz*cht!	Rinderzucht Rinderzuchtverband Rinderzüchter	*Cattle breeding* *Cattle breeding association* *Cattle breeder*
rogen!	rogenherstellung! Rogenproduktion Rogen	*Roe production* *Roe production* *Roe*
saibling!	Saiblingzucht Saiblingproduktion Saiblinge	*Char farming* *Char production* *Chars*
schweinehalt!	Schweinehaltung Schweinhalterinnen Schweinehalter	*Pig farming* *Pig farmers (female)* *Pig farmers*
schweinem*st!	Schweinemast Schweinemastbetrieb Schweinemastanlage	*Pig fattening* *Pig fattening operation* *Pig fattening facility*
schweineprod!	Schweineproduktion Schweineproduzent Schweineprodukte	*Pig production* *Pig producer* *Pig products*
schweinez*cht!	Schweinezucht Schweinezuchtverband Schweinezüchter	*Pig breeding* *Pig breeding association* *Pig breeder*

German term	Finds (maximum 3 examples each)	*Translation*
teichw!	Teichwirt Teichwirtschaft Teichwirtin	*Pond farmer* *Pond management* *Pond farmer (female)*
tierleid!	Tierleid Tierleidskandal -	*Animal suffering* *Animal suffering scandal* -
tierm*st!	Tiermast Tiermastbetrieb Tiermedizinspezialist	*Animal fattening* *Animal fattening operation* *Veterinary specialist*
tierproduktion!	Tierproduktion - -	*Animal production* - -
tierqu*l!	Tierquälerei Tierquälerisch Tierquäler	*Animal cruelty* *Cruel to animals* *Animal abuser*
tiersch*tz!	Tierschutz Tierschutzverstoß Tierschutzgesetz	*Animal welfare* *Animal welfare violation* *Animal welfare law*
tiertransport!	Tiertransport Tiertransporte Tiertransportgesetz	*Animal transport* *Animal transports* *Animal transport law*
tierversuch!	Tierversuch Tierversuchszweck Tierversuche	*Animal testing* *Animal testing purpose* *Animal tests*
tierz*cht!	Tierzucht Tierzuchtverband Tierzuchtbetrieb	*Animal breeding* *Animal breeding association* *Animal breeding operation*
versuchstier!	Versuchstier Versuchstiere Versuchstierhaltung	*Laboratory animal* *Laboratory animals* *Laboratory animal keeping*
zierv*gel!	Ziervögel Ziervogel Ziervogelhaltung	*Ornamental birds* *Ornamental bird* *Ornamental bird keeping*
zirkus!	Zirkustiere Zikuszelt Zirkus	*Circus animals* *Circus tent* *Circus*
zoo!	Zootiere Zootierhaltung Zoo	*Zoo animals* *Zoo animal keeping* *Zoo*
*forelle!	Forellenzucht Forellenteich Forellenkaviar	*Trout farming* *Trout pond* *Trout caviar*
*karpfen!	Karpfenteichwirtschaft Karpfenernte Koi-Karpfen	*Carp pond management* *Carp harvest* *Koi carp*
*kaviar!	Kaviarhandel Kaviarproduktion Kaviarbestände	*Caviar trade* *Caviar production* *Caviar stocks*

German term	Finds (maximum 3 examples each)	*Translation*
*muschel!	Muschelzucht Muschelfischerei Muschelkulturen	*Mussel farming* *Mussel fishing* *Mussel cultures*

Appendix B. Iterations of relevance criteria

Criteria iteration 1 (not reliable)

kr1: The text is relevant if at least one contextual relationship is established between the terms (tiersch*tz) and landw! or one of its synonyms.
kr2: The text is relevant if (kr1) does not apply, but other terms were used that are not yet part of the search term and nevertheless establish a relationship in the sense of the research question.

Even if (kr1) or (kr2) apply, the text is still not relevant if:

kr3: it is a table of contents, a bibliography or another type of list.
kr4: it concerns letters to the editor or opinions that do not originate from the editorial office or the journalistic environment.
kr5: it concerns legal texts or texts that present laws in a descriptive manner.
kr6: the terms are mentioned in passing, comparatively or enumeratively.

Criteria iteration 2 (not reliable)

kr1: The text is relevant if at least one contextual relationship is established between the terms (tiersch*tz) and landw! or one of its synonyms. The contextual relationship is irrelevant,
kr1.1: whether the context of the terms is food and food production, or another use of farm animals kept in agriculture.
kr1.2: whether the geographical context of the terms is Germany.
kr1.3: whether the temporal context is within the years 2010–2023.
kr2: The text is relevant if (kr1) does not apply, but other terms were used that are not yet part of the search term and nevertheless establish a relationship in the sense of the research question. “kr1.1” to “kr1.3” also apply to these terms.

Even if (kr1) or (kr2) apply, the text is still not relevant if:

kr3: it is a table of contents, a bibliography or another type of listing.
kr4: it concerns letters to the editor or opinions that do not originate from the editorial office or the journalistic environment.
kr5: they are legal texts or texts that present laws in a descriptive manner.
kr6: the terms are mentioned in passing, comparatively or enumeratively.
kr7: the contextual relationship has no agricultural context (e.g. animal husbandry and animal welfare in the context of zoos without an agricultural reference).

Criteria iteration 3 (reliable)

kr1: The text is relevant if at least one contextual relationship is established between the terms (tiersch*tz) AND landw! (OR one of its synonyms) (see search term A).
kr2: The text is relevant if it is an interview and the criterion kr1 applies.

kr3: The text is relevant if it is a press release and the criterion kr1 applies. With regard to the contextual relationship (see kr1), it is also relevant if:
skr1: in addition to or instead of the listed synonyms for landw! in search term A, the following terms from "aquaculture" refer to further synonyms for agricultural production and utilization systems for animals:
tiersch*tz **AND** aquakultur! OR fischw! OR teichw! OR fischz*cht! OR aal! *forelle! OR karpfen! OR *muschel! OR kaviar! OR rogen! OR saibling!
skr2: the context of the terms is a commercial use of farmed animals other than food production (e.g. production of fur).
skr3: the geographical context of the terms is outside Germany.
skr4: the temporal context of the terms is outside the years 2010–2023.

The text is not relevant if:

kr4: it is a contextual or bibliographical reference.
kr5: it concerns letters to the editor or opinions that do not originate from the editorial team or the journalistic environment.
kr6: they are legal texts, their uncommented reproduction or texts that present laws in a descriptive manner.
kr7: the terms of search term A (including the terms from Skr1) are mentioned in passing, comparatively or enumeratively and without contextual relationship.
kr8: the contextual relationship has no agricultural context (e.g. animal husbandry and animal welfare in the context of zoos or laboratory animals in research without an agricultural reference).
kr9: the agricultural context is fishing (fishing for wild fish in the sea) or hunting.

Appendix C. Variants of search terms

Open search term:

tiersch*tz **AND** (landw! OR agra! OR agrarier! OR agro! OR agronom! OR b*uer! OR geflügelhalt! OR geflügelm*st! OR legehenne! OR milchviehhalt! OR tierhalt! OR rinderhalt! OR rinderm*st! OR schweinehalt! OR schweinem*st!).

Search term B:

tiersch*tz **AND** (landw! OR agra! OR agrarier! OR agro! OR agronom! OR b*uer! OR tierhalt! OR tierproduktion! OR tiertransport! OR tierz*cht! OR tierm*st! OR geflügelhalt! OR geflügelm*st! OR geflügelz*cht! OR legehenne! OR küken! OR milchviehhalt! OR rinderhalt! OR rinderm*st! OR rinderz*cht! OR k*lb! OR schweinehalt! OR schweinem*st! OR ferkel! OR aquakultur! OR fischw! OR teichw! OR fischz*cht! OR aal! *forelle! OR karpfen! OR *muschel! OR kaviar! OR rogen! OR saibling!).

Seach term C:

(tierleid! OR tiersch*tz! OR tierqu*l!) **AND** (landw! OR agra! OR agrarier! OR agro! OR agronom! OR b*uer! OR nutztierhalt! OR tierproduktion! OR tiertransport! OR tierz*cht! OR tierm*st! OR geflügelhalt! OR geflügelm*st! OR geflügelz*cht! OR legehenne! OR küken! OR milchviehhalt! OR rinderhalt! OR rinderm*st! OR rinderz*cht! OR k*lb! OR schweinehalt! OR schweinem*st! OR ferkel! OR aquakultur! OR fischw! OR teichw! OR fischz*cht! OR aal! *forelle! OR karpfen! OR *muschel! OR kaviar! OR rogen! OR saibling!) **BUT NOT** (zoo! OR haustier! OR versuchstier! OR tierversuch! OR hund! OR katze! OR hamster! OR kleintier! OR zierv*gel! OR aquari! OR insekt!).

Search term D:

(tierleid! OR tiersch*tz! OR tierqu*l!) **AND** (landw! OR agra! OR agrarier! OR agro! OR agronom! OR b*uer! OR nutztierhalt! OR tierproduktion! OR tiertransport! OR tierz*cht! OR tierm*st! OR geflügelhalt! OR geflügelm*st! OR geflügelz*cht! OR geflügelprod! OR hähnchenhalt! OR hähnchenm*st! OR hähnchenz*cht! OR masth*hn! legehenne! OR legehennen! OR küken! OR milchviehhalt! OR milchviehz*cht! OR milchviehprod! OR rinderhalt! OR rinderm*st! OR rinderz*cht! OR rinderprod! OR k*lb! OR schweinehalt! OR schweinem*st! OR schweinez*cht! OR schweineprod! OR mastschwein! OR ferkel! OR aquakultur! OR fischw! OR teichw! OR fischz*cht! OR aal! *forelle! OR karpfen! OR *muschel! OR kaviar! OR rogen! OR saibling!) **BUT NOT** (zoo! OR zirkus! OR haustier! OR versuchstier! OR tierversuch! OR hund? OR katze! OR hamster! OR kleintier! OR zierv*gel! OR aquari!).

The final search term E:

(tierleid! OR tiersch*tz! OR tierqu*l!) **AND** (landw! OR agra! OR agrarier! OR agro! OR agronom! OR b?uer! OR nutztierhalt! OR tierproduktion! OR tiertransport! OR tierz*cht! OR tierm*st! OR geflügelhalt! OR geflügelm*st! OR geflügelz*cht! OR geflügelprod! OR hähnchenhalt! OR hähnchenm*st! OR hähnchenz*cht! OR masth*hn! OR h?hn! OR legehenne! OR legehennen! OR küken! OR milchviehhalt! OR milchviehz*cht! OR milchviehprod! OR rinderhalt! OR rinderm*st! OR rinderz*cht! OR rinderprod! OR k*lb! OR schweinehalt! OR schweinem*st! OR schweinez*cht! OR schweineprod! OR mastschwein! OR ferkel! OR aquakultur! OR fischw! OR teichw! OR fischz*cht! OR aal! *forelle! OR karpfen! OR *muschel! OR kaviar! OR rogen! OR saibling!) **BUT NOT** (zoo! OR zirkus! OR haustier! OR versuchstier! OR tierversuch! OR hund? OR katze! OR hamster! OR kleintier! OR zierv*gel! OR aquari! OR leser*brief!

Intuitive search term:

„tiersch*tz!‘ AND „landwirt!‘

AI-Assisted search term by ChatGPT (4.0)

((nutz*tier! OR massentier! OR intensivtier! OR tierhaltung! OR tiermast! OR masttier! OR fleischproduktion! OR stallhaltung! OR agrarindustrie! OR tierfabriken! OR fleischindustrie!) AND (schwein* OR rind* OR geflügel* OR huhn* OR pute* OR kalb* OR kuh* OR ochs*)) AND (tiersch*tz! OR tiersch*tzgesetz! OR tiersch*tzverstoß! OR tiersch*tzorganisation! OR tiersch*tzproblem! OR tierwohl! OR tierleid! OR tiersch*tzskandal! OR ethik’) BUT NOT (haustier! OR zootier! OR wildtier!).

AI-Assisted search term by Gemini (1.5 Flash model)

(intensiv* Nutztierhalt* OR Massentierhalt* OR Stallhalt* OR industrielle Landwirtschaft) AND (Tiersch*tz! OR Tierwohl OR Tierquälerei OR Tierschutzgesetz OR Verstoß* gegen Tierschutzgesetz).

Appendix D.

Table A1 is a closer look at the individual refinement steps that underlie the gradual performance improvement of the search term candidates (see details of the search terms in Appendix C). With $p = 0.632$, $r = 0.902$, and $F1 = 0.743$ (cf. Table 3), search term B did not show satisfactory performance. While recall r was good, precision p was too low. This was due to further terms for animal production that were added to the existing synonyms

Table A1. Stepwise refinements of the open search term[a].

Full version of the open search term (A): tiersch*tz AND (landw! OR agra! OR agrarier! OR agro! OR agronom! OR b*uer! OR geflügelhalt! OR geflügelm*st! OR legehenne! OR milchviehhalt! OR tierhalt! OR rinderhalt! OR rinderm*st! OR schweinehalt! OR schweinem*st!).

Variants of search terms	**Refinements to previous search term**
Search term B	added: *OR tierproduktion! OR tiertransport! OR tierz*cht! OR tierm*st! OR geflügelz*cht! OR küken! OR rinderz*cht! OR k*lb! OR ferkel! OR aquakultur! OR fischw! OR teichw! OR fischz*cht! OR aal! *forelle! OR karpfen! OR *muschel! OR kaviar! OR rogen! OR saibling!*
Search term C	– added: *tierleid! OR tierqu*l!* – *!* after *tiersch*tz* – *tierhalt!* replaced by *nutztiehalt!* – added: *BUT NOT (zoo! OR haustier! OR versuchstier! OR tierversuch! OR hund! OR katze! OR hamster! OR kleintier! OR zierv*gel! OR aquari! OR insekt!)*
Search term D	– added: *OR geflügelprod! OR hähnchenhalt! OR hähnchenm*st! OR hähnchenz*cht! OR masth*hn! OR legehennen! OR milchviehz*cht! OR milchviehprod! OR rinderprod! OR schweinez*cht! OR schweineprod! OR mastschwein!* – *insekt!* readmitted (deleted in the BUT NOT term) – *zirkus!* added to the BUT NOT term – *!* in *hund!* replaced by?
Search term E	– * in *b*uer!* replaced by? – * in *h*hn* replaced by? – *leser*brief!* added to the BUT NOT term

Full version of the final search term (E): (tierleid! OR tiersch*tz! OR tierqu*l!) **AND** (landw! OR agra! OR agrarier! OR agro! OR agronom! OR b?uer! OR nutztierhalt! OR tierproduktion! OR tiertransport! OR tierz*cht! OR tierm*st! OR geflügelhalt! OR geflügelm*st! OR geflügelz*cht! OR geflügelprod! OR hähnchenhalt! OR hähnchenm*st! OR hähnchenz*cht! OR masth*hn! OR h?hn! OR legehenne! OR legends! OR küken! OR milchviehhalt! OR milchviehz*cht! OR milchviehprod! OR rinderhalt! OR rinderm*st! OR rinderz*cht! OR rinderprod! OR k*lb! OR schweinehalt! OR schweinem*st! OR schweinez*cht! OR schweineprod! OR mastschwein! OR ferkel! OR aquakultur! OR fischw! OR teichw! OR fischz*cht! OR aal! *forelle! OR karpfen! OR *muschel! OR kaviar! OR rogen! OR saibling!) **BUT NOT** (zoo! OR zirkus! OR haustier! OR versuchstier! OR tierversuch! OR hund? OR katze! OR hamster! OR kleintier! OR zierv*gel! OR aquari! OR leser*brief!)

[a] A brief description of the terms contained in the search terms A–E is provided in English in Appendix A.

for farmer and agriculture (landw*) in the open search term. The additional terms limited the animal species to those with Germany's largest livestock populations (Statista 2024) and those that are most frequently produced in aquaculture farms. Thus, search term B found almost all of the articles, both relevant and nonrelevant, that were also found by the open search term.

After some more modifications, we obtained search term C, with an increase in precision to $p = 0.774$ (+0.141), but also a decrease in recall to $r = 0.672$ (-0.230) and a decrease in their harmonic mean to $F1 = 0.719$ (-0.024). The changes from B to C included additional terms that resulted in articles being excluded if they contained at least one of these terms. As shown in the Table above, the OR-linked terms were appended to the existing search term in brackets with the connector "BUT NOT". These terms were considered indicators of non-relevant articles and were also obtained from the review process when the relevance criteria were established. These included terms related to animal welfare and animal husbandry but without reference to agriculture (e.g. zoo or pets). However, the effect of the exclusionary terms was so strong that only 41 of the 61 relevant articles were identified. Consequently, the results for r and $F1$ were the lowest of all the versions.

The performance of search term D, with $p = 0.779$ (+0.006) and $r = 0.869$ (+0.197), was quite good and very close to the final search term E. But, while the summary performance index of term D was $F1 = 0.852$ (+0.102), the stopping rule was not reached, as precision was still below 0.8. The increase in performance from C to D was obtained by adding terms that complemented the existing animal husbandry terms for animal species, breeding, and production areas. The term "insects!" was removed from the term "BUT NOT" because many relevant articles also mentioned insects—for example, when discussing alternatives to food production in animal husbandry. These articles were erroneously excluded by the previous search term. It was also necessary to replace the ! in "Hund!" with a ?, as otherwise not only terms such as "Hund" (relates to "dog") or "Hunde" (relates to "dogs") would be excluded but also terms like "Hunderte" (relates to a "hundreds"), causing relevant articles to be missed.

After further adjustments, we reached the stopping rule and obtained the final search term E, with $p = 0.825$ (+0.046), $r = 0.852$ (−0.016), and $F1 = 0.839$ (+0.017). The decisive improvement in the performance of search term E was achieved through various refinements of the terms "b*uer" and "h*hn". For example, "b*uer!" also included texts in thehit list if "Berufsfeuerwehr" (relates to "professional fire department") appeared. To find only terms with the word stem Bauer (relates to "farmer"), the * in the search term D "b*uer!" had to be replaced with a ?. The same was done for the term "h*hn!" to find "Huhn", "Hahn" or "Hähnchen" (relates to "poultry") but not "Herkunftsbezeichnung" (relates to "designation of origin").

References

Baker M. (2016) '1,500 scientists Lift the Lid on Reproducibility', *Nature*, 533/7604:452–4. https://doi.org/10.1038/533452a

Baker S. R., Bloom N. and Davis S. J. (2016) 'Measuring Economic Policy Uncertainty', *The Quarterly Journal of Economics*, 131/4:1593–636. https://doi.org/10.1093/qje/qjw024

Barberá P. et al. (2021) 'Automated Text Classification of News Articles: a Practical Guide', *Political Analysis*, 29/1:19–42. https://doi.org/10.1017/pan.2020.8

Benoit K. (2020) 'Text as Data: an Overview', In: *The SAGE Handbook of Research Methods in Political Science and International Relations*. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications Ltd. https://doi.org/10.4135/9781526486387

Bonfadelli H. and Friemel T. N. (2017) *Medienwirkungsforschung*, 6, überarbeitete Auflage. Konstanz: UVK Verlagsgesellschaft mbH.

Boumans J. W. and 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

Entman R. M. (1993) 'Framing: toward Clarification of a Fractured Paradigm', *Journal of Communication*, 43/4:51–8. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Gentzkow M., Kelly B. and Taddy M. (2019) 'Text as Data', *Journal of Economic Literature*, 57/3:535–74. https://doi.org/10.1257/jel.20181020

Grimmer J. and Stewart B. M. (2013) 'Text as Data: the Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts', *Political Analysis*, 21/3:267–97. https://doi.org/10.1093/pan/mps028

Hassan T. A. et al. (2019) 'Firm-Level Political Risk: Measurement and Effects*', *The Quarterly Journal of Economics*, 134/4:2135–202. https://doi.org/10.1093/qje/qjz021

Hayes A. F. and Krippendorff K. (2007) 'Answering the Call for a Standard Reliability Measure for Coding Data', *Communication Methods and Measures*, 1/1:77–89. https://doi.org/10.1080/19312450709336664

IVW. (2022) 'Ranking Der Auflagenstärksten Überregionalen Tageszeitungen in Deutschland im 3. Quartal 2022'. *Statista*, https://de.statista.com/statistik/daten/studie/73448/umfrage/auflage-der-ueberregionalen-tageszeitungen/ Retrieved 11 January 2023.

Janker J. and Mann S. (2020) 'Understanding the Social Dimension of Sustainability in Agriculture: a Critical Review of Sustainability Assessment Tools', *Environment, Development and Sustainability*, 22/3:1671–91. https://doi.org/10.1007/s10668-018-0282-0

Jarren O. and Donges P. (2011) *Politische Kommunikation in der Mediengesellschaft*. Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-93446-4

Jellason N. P. et al. (2022) 'A Systematic Review of Smallholder Farmers' climate Change Adaptation and Enabling Conditions for Knowledge Integration in Sub-Saharan African (SSA) Drylands', *Environmental Development*, 43:100733. https://doi.org/10.1016/j.envdev.2022.100733

Jünger J. and Geise S. (2022) 'Unboxing Computational Social Media Research from a Datahermeneutical Perspective: How Do Scholars Address the Tension between Automation and Interpretation?', *International Journal of Communication*, 16: 1482–1505.

King G., Lam P. and Roberts M. E. (2017) 'Computer-Assisted Keyword and Document Set Discovery from Unstructured Text', *American Journal of Political Science*, 61/4:971–88. https://doi.org/10.1111/ajps.12291

Kolbe R. H. and Burnett M. S. (1991) 'Content-Analysis Research: an Examination of Applications with Directives for Improving Research Reliability and Objectivity', *Journal of Consumer Research*, 18/2:243. https://doi.org/10.1086/209256

Krippendorff K. (2004a) *Content Analysis: an Introduction to Its Methodology*. 2nd ed. Thousand Oaks, CA: Sage.

——— (2004b) 'Reliability in Content Analysis.: some Common Misconceptions and Recommendations', *Human Communication Research*, 30/3:411–33. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Kroon A. C., van der Meer T. and Vliegenthart R. (2022) 'Beyond Counting Words: Assessing Performance of Dictionaries, Supervised Machine Learning, and Embeddings in Topic and Frame Classification', *Computational Communication Research*, 4/2:528–70. https://doi.org/10.5117/CCR2022.2.006.KROO

Lacy S. et al. (2015) 'Issues and Best Practices in Content Analysis', *Journalism & Mass Communication Quarterly*, 92/4:791–811.

Lehberger M. and Gruener S. (2023) "(Why) Do Farmers' Big Five Personality Traits Matter?—A Systematic Literature Review'. https://doi.org/10.31219/osf.io/jbx4p

Lombard M., Snyder-Duch J. and Bracken C. C. (2002) 'Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability', *Human Communication Research*, 28/4:587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Mahl D., von Nordheim G. and Guenther L. (2022) 'Noise Pollution: a Multi-Step Approach to Assessing the Consequences of (Not) Validating Search Terms on Automated Content Analyses', *Digital Journalism*, 2:1–23. https://doi.org/10.1080/21670811.2022.2114920

Maier D. et al. (2020) 'How Document Sampling and Vocabulary Pruning Affect the Results of Topic Models', *Computational Communication Research*, 2/2:139–52. https://doi.org/10.5117/CCR2020.2.001.MAIE

Matthes J. and Kohring M. (2008) 'The Content Analysis of Media Frames: toward Improving Reliability and Validity', *Journal of Communication*, 58/2:258–79. https://doi.org/10.1111/j.1460-2466.2008.00384.x

Mizik T. (2023) 'How Can Precision Farming Work on a Small Scale? A Systematic Literature Review', *Precision Agriculture*, 24/1:384–406. https://doi.org/10.1007/s11119-022-09934-y

Mohr S. and Höhler J. (2023) 'Media Coverage of Digitalization in Agriculture—an Analysis of media Content', *Technological Forecasting and Social Change*, 187:122238. https://doi.org/10.1016/j.techfore.2022.122238

Müller H. et al. (2018) 'Der Wert der Worte—Wie Digitale Methoden Helfen, Kommunikations- und Wirtschaftswissenschaft Zu Verknüpfen', *Publizistik*, 63/4:557–82. https://doi.org/10.1007/s11616-018-0461-x

Neuendorf K. A. (2017) *The Content Analysis Guidebook*. Thousand Oaks, CA: SAGE Publications, Inc. https://doi.org/10.4135/9781071802878

Nosek B. A. et al. (2022) 'Replicability, Robustness, and Reproducibility in Psychological Science', *Annual Review of Psychology*, 73/1: 719–48. https://doi.org/10.1146/annurev-psych-020821-114157

O'Connor C. and Joffe H. (2020) 'Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines', *International Journal of Qualitative Methods*, 19:160940691989922. https://doi.org/10.1177/1609406919899220

Riffe D. et al. (2019) *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. 4th ed. New York: Routledge. https://doi.org/10.4324/9780429464287

Rommel J. et al. (2023) 'Farmers' risk Preferences in 11 European Farming Systems: a Multi-country Replication of Bocquého et al. (2014)', *Applied Economic Perspectives and Policy*, 45/3:1374–99. https://doi.org/10.1002/aepp.13330

Scheufele B. and Engelmann I. (2016) 'Journalismus und Framing'. in: M., Löffelholz, L., Rothenberger (eds.) *Handbuch Journalismustheorien*, pp. 443–56. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-531-18966-6_27

Schulze C. et al. (2024) 'Using Farmers' ex Ante Preferences to Design Agri-environmental Contracts: a Systematic Review', *Journal of Agricultural Economics*, 75/1:44–83. https://doi.org/10.1111/1477-9552.12570

Shiller R. J. (2019) *Narrative Economics, How Stories Go Viral and Drive Major Economic Events*. Princeton, NJ: PUP. https://doi.org/10.1515/9780691189970

Statista. (2024) 'Nutztierbestand in Deutschland 1900-2023'. *Statista*. https://de.statista.com/statistik/daten/studie/659045/umfrage/nutztierbestand-in-deutschland/ Retrieved 6 May 2024.

Stryker J. E. et al. (2006) 'Validation of Database Search Terms for Content Analysis: the Case of Cancer News Coverage', *Journalism & Mass Communication Quarterly*, 83/2:413–30.

Thompson B. et al. (2023) 'Farmers' adoption of Ecological Practices: a Systematic Literature Map', *Journal of Agricultural Economics*, 75:84–107. https://doi.org/10.1111/1477-9552.12545

Van Atteveldt W., Van Der Velden M. A. C. G. and Boukes M. (2021) 'The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms', *Communication Methods and Measures*, 15/2:121–40. https://doi.org/10.1080/19312458.2020.1869198

Vecchio R. and Cavallo C. (2019) 'Increasing Healthy Food Choices through Nudges: a Systematic Review', *Food Quality and Preference*, 78:103714. https://doi.org/10.1016/j.foodqual.2019.05.014

Velten S. et al. (2015) 'What Is Sustainable Agriculture? A Systematic Review', *Sustainability*, 7/6: 7833–65. https://doi.org/10.3390/su7067833

Vindigni G. et al. (2021) 'Shedding Light on Peri-Urban Ecosystem Services Using Automated Content Analysis', *Sustainability*, 13/16:9182. https://doi.org/10.3390/su13169182