COMMUNICATION METHODS AND MEASURES
2022, VOL. 16, NO. 4, 303–320
https://doi.org/10.1080/19312458.2021.1999913

Routledge
Taylor & Francis Group

Enhancing Theory-Informed Dictionary Approaches with "Glass-box" Machine Learning: The Case of Integrative Complexity in Social Media Comments

Timo Dobbrick[a], Julia Jakob[a], Chung-Hong Chan[a], and Hartmut Wessler[b]

[a]Mannheim Centre for European Social Research, University of Mannheim; [b]Institute for Media and Communication Studies, University of Mannheim

ABSTRACT
Dictionary-based approaches to computational text analysis have been shown to perform relatively poorly, particularly when the dictionaries rely on simple bags of words, are not specified for the domain under study, and add word scores without weighting. While machine learning approaches usually perform better, they offer little insight into (a) which of the assumptions underlying dictionary approaches (bag-of-words, domain transferability, or additivity) impedes performance most, and (b) which language features drive the algorithmic classification most strongly. To fill both gaps, we offer a systematic assumption-based error analysis, using the integrative complexity of social media comments as our case in point. We show that attacking the additivity assumption offers the strongest potential for improving dictionary performance. We also propose to combine off-the-shelf dictionaries with supervised "glass box" machine learning algorithms (as opposed to the usual "black box" machine learning approaches) to classify texts and learn about the most important features for classification. This dictionary-plus-supervised-learning approach performs similarly well as classic full-text machine learning or deep learning approaches, but yields interpretable results in addition, which can inform theory development on top of enabling a valid classification.

When analyzing large amounts of digital data, various automated measurement tools are available to social scientists (Boumans & Trilling, 2015). In the past, the field has relied heavily on off-the-shelf dictionaries that search for a pre-defined list of words to parse such corpora (Dun et al., 2020). More recently, however, there has been a shift toward using more enhanced supervised and unsupervised machine learning approaches for this task, because they regularly outperform dictionary-based text classifications (Van Atteveldt et al., 2021).

Off-the-shelf dictionaries are widely appreciated for their simplicity. They are easy to implement and just as easy to interpret, in addition to being rather cost-efficient since they require little human input (González-Bailón & Paltoglou, 2015). This makes them attractive to a broad range of researchers, including those that received traditional as opposed to computational methodological training (Boumans & Trilling, 2015). However, major challenges of off-the-shelf dictionaries evolve around the problems arising from the bag-of-words, domain transferability, and additivity assumptions that are underlying this approach (Chan et al., 2021).

Some scholars seem to implicitly assume that off-the-shelf dictionaries can be readily transferred from one research context to another while producing similarly valid and reliable classifications (domain transferability assumption). However, studies show that the validity of off-the-shelf

CONTACT Timo Dobbrick Timo.Dobbrick@mzes.uni-mannheim.de Mannheimer Zentrum für Europäische Sozialforschung (MZES), A 5, 6, A - Raum 215, Mannheim, 68159

dictionaries varies widely across different datasets (Ribeiro et al., 2016) and that "their performance is poorer for less diverse, domain specific content" (González-Bailón & Paltoglou, 2015, p. 105). This validity problem can be alleviated by manually validating off-the-shelf tools or by creating tailor-made dictionaries, which have been shown to work well in the context for which they were developed (Haselmayer & Jenny, 2017; Muddiman et al., 2019; Young & Soroka, 2012). Still, both off-the-shelf and context-tailored dictionaries perform rather poorly in relation to human coding and more sophisticated automated methods, with low reliabilities that are barely better than chance (Boukes et al., 2020; Van Atteveldt et al., 2021). This is because dictionaries implement a bag-of-words approach, which presumes that the words in a statement are semantically independent (bag-of-words assumption) (Young & Soroka, 2012). Hence, dictionaries disregard the order of words in a sentence and thereby the grammatical structure of texts (Chan et al., 2021). Moreover, word lists usually operate on an additivity assumption, meaning that all dictionary features are equally relevant to the presence of a sought-after construct (Young & Soroka, 2012). Thus, statements that include a higher number of dictionary words often score stronger on a specific construct, that is, for example, a message that contains the words "good" and "excellent" would be considered more positive than one that features either the one or the other expression (Chan et al., 2021). This neglects that "in natural language, of course, certain words may carry more weight than others" (Young & Soroka, 2012, p. 209), like for example, "excellent" could be more relevant than "good" when measuring positive sentiment. A key question in this context is how to determine the weight that should be assigned to each dictionary feature (Young & Soroka, 2012). Furthermore, the additivity assumption again disregards grammatical specifics, for instance, by ignoring adverbs that amplify certain words (e.g., "good" vs. "very good") (Chan et al., 2021).

Supervised and unsupervised machine learning approaches can solve these problems and reach significantly better agreements with manually coded data (González-Bailón & Paltoglou, 2015; Rudkowsky et al., 2018; Stoll et al., 2020; Van Atteveldt et al., 2021). However, as most research in this area implements a shotgun approach, to date, there is little insight into *why* machine learning works better than dictionary-based methods. Since extant studies usually address the bag-of-words, domain transferability, and additivity assumptions all at once, it is ultimately unclear which of these problems is tackled by the machine learning.

The present study aims to address this gap. It follows an assumption-based framework that tackles the bag-of-words, domain transferability, and additivity assumption one at a time to locate the major source of weakness in off-the-shelf dictionary-based methods. This multi-step error analysis (Ng, 2018) shows which assumption needs to be addressed for the largest improvement in agreement with hand-coded gold standards. The study then also compares this assumption-based approach against the performance of full-text machine learning methods to demonstrate the advantages of "dictionary-plus-supervised-learning" combinations that have recently been suggested by Dun et al. (2020). The dictionary-plus-supervised-learning approach we implement here tackles the additivity assumption by using supervised machine learning to improve and extend the dictionary-based classification. Specifically, this means that the machine learning algorithm is trained on a hand-coded data set that has previously been pre-processed by using an off-the-shelf dictionary, instead of using the full text items. The automated classifier thus capitalizes on the information that is contained in the dictionary categories. Two main advantages of this approach are that it can reduce the quantity of hand-coding that is needed for full-text machine learning and that it provides theoretically interpretable insights into which dictionary features prove important in the automated classification.

Our main goal in this study is not to criticize off-the-shelf dictionaries as others have done before us, but to diagnose the primary origin of their weakness and to show how researchers can nevertheless leverage the theoretical information that is carried by pre-defined word lists. Computational social science is at a point now where a closer link should be forged between conceptual reflections and automated measurements, and where researchers need to move beyond binary classifications toward the detection of more sophisticated theory-driven constructs (Baden et al., 2020). By means of automatically classifying an ordinally hand-coded content-related measure, namely the integrative complexity, this paper uncovers the potential held for this task by combining off-the-shelf dictionaries with supervised machine learning.

The case study: Integrative complexity

Our interest in developing state-of-the-art automated text analysis methods largely stems from our goal to study the quality of user-generated political discussions on a large scale across different platforms and countries. Therefore, our case study in this paper is to detect the integrative complexity of English- and German-language online user comments from Facebook, Twitter, and news website comment sections. In so doing, we add to the existing literature not only by locating the major source of weakness in dictionary-based methods, but also by testing the applicability of different automated text analysis tools in assessing complex non-sentiment-based constructs and exploring the potential of these tools for comparative research that involves non-English languages.

Integrative complexity is a psychological measure that researchers increasingly implement to assess the argumentative quality of public debate contributions (e.g., Moore et al., 2020; Wyss et al., 2015). It captures the sophistication of written statements by assessing their degree of differentiation and integration as two consecutive dimensions of complexity (Suedfeld et al., 1992). While differentiated user comments consider several aspects or viewpoints with respect to a given topic, integration occurs when these are linked to each other on a more abstract level, for example, by pinpointing shared attributes, superordinate principles or conflicting goals (Moore et al., 2020).

The manual coding scale for integrative complexity is ordinal and ranges from 1 (lowest complexity) to 7 (highest complexity) (Baker-Brown et al., 1992). Simplistic user comments that lack both differentiation and integration receive a score of 1, whereas contributions that consider two or more aspects of or perspectives on an issue but fail to integrate these are given a score of 3. Posts that contain moderately or strongly developed conceptual integrations are high in complexity and obtain a score of 5 or 7, respectively. The scores of 2, 4, and 6 mirror transitional stages of complexity where "differentiation and integration are implicit and emergent rather than explicit and fully articulated" (Baker-Brown et al., 1992, p. 402).

The dominant approach to automatically score integrative complexity was made popular in the social sciences by Wyss et al. (2015) and is to combine ten distinct categories from the off-the-shelf Linguistic Inquiry and Word Count (LIWC) dictionary (see also Beste & Wyss, 2014; Kesting et al., 2018; Moore et al., 2020). LIWC is widely used for automated content and sentiment analysis (Pennebaker et al., 2015). Apart from linguistic, psychological, and personal dimensions based on dictionary words, it also includes formal features such as word count (WC) or the average number of words per sentence (WPS) (Pennebaker, et al., 2007). The relevant LIWC dimensions for measuring integrative complexity have been derived theoretically by Owens and Wedeking (2011, pp. 1055–1057) and include:

(1) *The percentage of words in a text with six or more letters* – a higher share of six-letter words is associated with communicative sophistication and thus increased integrative complexity (*Sixl*)
(2) *Discrepancy* – words like "should", "would", or "could" that indicate whether a statement points to differences or inconsistencies between multiple elements or perspectives, the identification of which increases integrative complexity (*Discr*)
(3) *Tentativeness* – words like "maybe", "perhaps", or "guess" that signal an openness to other arguments and perspectives on behalf of the author, which is connected to the reflection and consideration of new or alternative interpretations and thus a higher integrative complexity (*Tent*)

(4) *Inclusiveness* – words like “and”, “with”, or “include” that are markers for the extent to which a statement identifies relationships and connections between multiple differentiated aspects or perspectives, which raises integrative complexity (*Incl*)
(5) *Causation* – words like “because”, “effect”, or “hence” that reflect the identification of causal connections and linkages among different elements in a statement and thus increase integrative complexity (*Cause*)
(6) *Insight* – words like “think”, “know”, or “consider” that signal an advanced depth of understanding on behalf of an author, whereby a higher degree of insight into an issue corresponds to higher integrative complexity (*Insig*)
(7) *Inhibition* – words like “block”, “constrain”, or “stop” that suggest a certain degree of restraint is exercised by the author of a comment, which indicates their ability to reflect on how their actions or ideas are constrained, and is associated with greater integrative complexity (*Inhib*)
(8) *Certainty* – words like “always”, or “never” that, in contrast to tentativeness or inhibition markers, are typical for simple and undifferentiated utterances in which the author accepts only one possible interpretation of events, which decreases integrative complexity (*Cert*)
(9) *Negations* – words like “no”, “not”, or “never” that are indicative of one-dimensional and absolute rule structures; they signal that the author of a statement accepts only one (often highly evaluative) view of the world and only one legitimate perspective on an issue with no room for alternative arguments or interpretations, which is associated with lower integrative complexity (*Negate*)
(10) *Exclusiveness* – words like “but”, “without”, or “exclude” that indicate whether an author distinctly separates their ideas from other perspectives and, in this context, are often used to exclude other interpretations from being valid, which lowers integrative complexity (*Excl*)

Following these theoretical reflections, the integrative complexity score is obtained by using the ten LIWC categories as a formula (Wyss et al., 2015). Below, we use this operationalization as a baseline model that we compare to the multiple steps of our assumption-based error analysis framework and the shotgun full-text machine learning approaches. The formula reads:

$$\mathrm{IC} = \mathrm{Sixl} + \mathrm{Discr} + \mathrm{Tent} + \mathrm{Incl} + \mathrm{Cause} + \mathrm{Insig} + \mathrm{Inhib} - \mathrm{Cert} - \mathrm{Negate} - \mathrm{Excl} \qquad (1)$$

Methodology

Data set

In contrast to studies that rely on one type of data from a specific national context, we draw on a gold standard sample that consists of a diverse set of 4,800 online user comments from different communication arenas and countries. These stem from the a) website comment sections, b) public Facebook pages of mainstream news media, from the c) Facebook pages of partisan collective actors and alternative media, and d) from Twitter, as well as from Australia, the United States, Germany, and Switzerland. Our study thus strengthens previous conclusions on the performance of various automated text classification approaches by footing them on a broader basis in terms of both platforms and countries (Van Atteveldt et al., 2021). The 4,800 gold standard comments were randomly sampled from a data set of N = 1,236,551 collected user contributions on the public role of religion and secularism in society (for further information on the data collection see Online Appendix I). We drew 300 comments from each of the four communication arenas in each of the four countries. The sample thus contains an equal number of comments from English- and German-speaking countries. Please note that even though Switzerland is a multi-lingual country, we focused on German-speaking discourses only to keep the number of analyzed languages manageable.

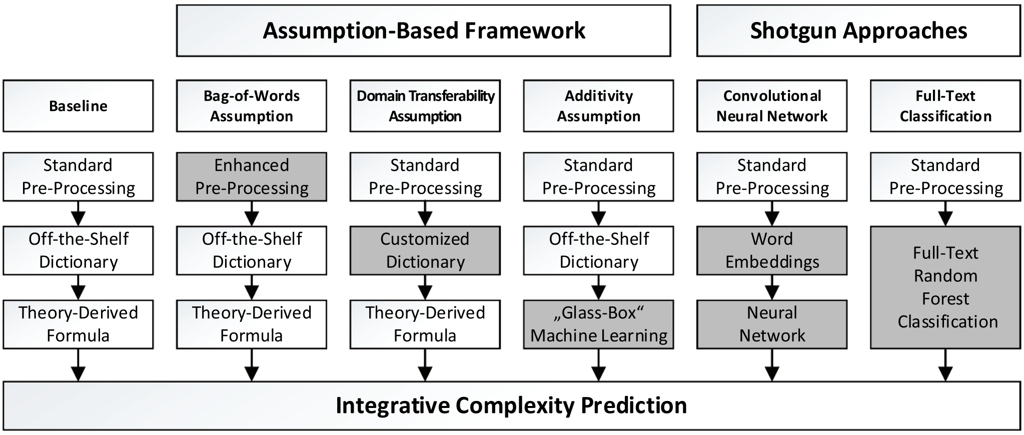

Figure 1. Overview of the Baseline, Assumption-Based Framework, and Shotgun Approaches. *Note*: Boxes highlighted in gray show the process steps that differ from the baseline process.

Gold standard coding

The gold standard was coded by three trained human coders based on the previously described scale from 1 (lowest complexity) to 7 (highest complexity) (Baker-Brown et al., 1992). In a pretest with 100 stratified randomly selected user comments, the three raters reached a Krippendorff's α_{ordinal} reliability of .85. Each gold standard comment was then coded by two individuals, with disagreements being resolved consensually. The coder pairs had reached Krippendorff's α_{ordinal} reliability of .88 and .86 in the pretest, respectively.

Assumption-based approach versus shotgun approaches

We propose an assumption-based framework to diagnose the major source of weakness in off-the-shelf dictionary-based classifications. Thereby, we address the three assumptions of dictionary approaches (i.e., bag-of-words, domain transferability, and additivity) one at a time using methods suggested in the literature. Based on the concept of error analysis, we evaluate tackling which assumption brings the biggest improvement in agreement between the modified analysis and the hand-coded gold standard. We then pit against the assumption-based approach with a shotgun full-text machine learning approach, in which multiple assumptions are tackled at the same time. Figure 1 gives an overview over these approaches.

Baseline: Integrative complexity score

The baseline model is the theory-derived, dictionary-based integrative complexity (IC) score by Wyss et al. (2015).[1] The original model was proposed to analyze argumentative complexity in Swiss parliamentary debates. As elaborated previously, the score is based on the LIWC dictionary and operationalized as follows:

$$\text{IC} = \text{Sixl} + \text{Discr} + \text{Tent} + \text{Incl} + \text{Cause} + \text{Insig} + \text{Inhib} - \text{Cert} - \text{Negate} - \text{Excl} \qquad (2)$$

[1] We recognize that there are several ways to capture textual complexity, and that syntax-based readability measures such as the Flesch-Kincaid score (Kincaid et al.,
) can also be used for this purpose. As these readability scores are based on formal elements such as the number of words, syllables,

This formula can be reasoned as the summation of every LIWC score with a weight coefficient. The first seven terms have a +1 weight coefficient, whilst the last three terms have a −1 weight coefficient. All the other terms from the LIWC dictionary have a 0 weight coefficient. A restatement of the IC formula is:

$$\begin{aligned} IC = {} & 1^{*}Sixl + 1^{*}Discr + 1^{*}Tent + 1^{*}Incl + 1^{*}Cause + 1^{*}Insig + 1^{*}Inhib \\ & +(-1)^{*}Cert+(-1)^{*}Negate+(-1)^{*}Excl + 0^{*}WC + 0^{*}Analytic + 0^{*} \\ & Clout+\ldots \end{aligned} \tag{3}$$

In Wyss et al. (2015), the score was calculated for French and German speech acts, for which the English LIWC2007 dictionary was translated into these languages by professional translators. In this study, we used the original English LIWC2007 dictionary for pre-processing our English-speaking gold standard comments. The German-speaking comments were pre-processed with the translated German LIWC dictionary by Wyss et al. (2015) so that our results are comparable to the original baseline IC score.

Assumption-based approach

As stated above, we attacked the three assumptions that underlie this dictionary approach one at a time to see tackling which assumption can generate the greatest agreement with the human-coded gold standard. In all cases, we divided the data into 10 equal parts and computed the accuracy metrics for each part. The overall accuracy metric is the average value of the 10 values. For the case of domain transferability and additivity, 10-fold cross validation was used, i.e., 9 parts were used for training and 1 part was used for validation.

Bag-of-words assumption

The bag-of-words assumption states that "the words people use convey psychological information over and above their literal meaning and independent of their semantic context" (Pennebaker et al., 2003, p. 550). Under this assumption, the word order and grammatical functions of words are not considered. For instance, "my cat is bad" and "is not my cat bad?" might be processed as having the same meaning, if the dictionary does not consider the grammatical function of the word "not" as well as the difference in sentence structure. Most of the academic attention on modifying text analysis was on attacking this assumption. Usually, the proposed solutions to tackle the bag-of-words assumption are in the form of another shotgun approach, i.e., they address different assumptions about word order and grammatical functions without knowing which of these problems is actually being addressed. Instead, we isolate components that can be tested independently. Some suggestions such as n-grams or word embeddings are not tested because they are dealing with more than one assumption. When using n-grams and word embeddings, the weight coefficients need to be adjusted as well and we cannot use the IC formula.

Negation. We followed Young and Soroka (2012) and paid special attention to words after negation signifiers such as "not," "no," "never," "hardly," and "less." In Young and Soroka (2012), words after these negation signifiers were considered to have the opposite meaning. For example, the positive word "good" is considered to have a negative meaning if it is preceded by "hardly," i.e., "hardly good" is considered negative. Unlike sentiment analysis, many categories in LIWC – such as words with six or more letters – do not have a complementary category with the flipped meaning. Instead, we followed Haselmayer and Jenny (2017) to exclude all words located after these negation signifiers.

and sentences, they are no dictionary-based methods and therefore not used as a baseline in this study. However, an additional baseline analysis that uses the Flesch-Kincaid score is available in the online appendix. It shows a very weak correlation of the Flesch-Kincaid score from both the English and German data with the respective hand-coded integrative complexity score.

Part of speech. Part-of-speech (POS) tagging refers to the process of marking up the grammatical function of a word in a sentence by considering both the meaning of the word and the context of the word being used. For example, "arms" in the sentence "There was a clatter as the basilisk fangs cascaded out of Hermione's arms" should be tagged as common noun, but should be tagged as a verb in the sentence "What sort of imbecile arms an assassin with his own blade?." In our framework, we tagged the POS of all words in our corpus using spacyr (Honnibal et al., 2020; Kenneth & Matsuo, 2020) and then applied the LIWC dictionary. We followed previous works to select only words with specific grammatical functions. Jacobi et al. (2015) selected only common nouns, proper nouns, verbs, and adjectives. Benamara et al. (2007) selected only adjectives and adverbs. We tried both in our framework.

Lemmatization. Lemmatization refers to the process of reducing a word to its fundamental form. For example, the plural "analyses" is reduced to the singular form "analysis." This procedure is more important for languages with a diverse set of conjugation rules, e.g., German or Hindi. How to lemmatize a word depends on the grammatical functions of the word in the specific context. Using the two sentences with the word "arms" above, the first instance of "arms" would be lemmatized to "arm" (a limb), whilst the second instance would be lemmatized to "to arm" (to pick up or supply weapons). In our framework, we lemmatized all words using spacyr (Honnibal et al., 2020; Kenneth & Matsuo, 2020) and then applied the LIWC dictionary.

Domain transferability assumption

We modified the Lexicon Expansion and Reduction methods by Diesner and Evans (2015). The original method proposes to rank terms in the entire corpus by TF-IDF, and then terms with a high TF-IDF value are manually evaluated for either inclusion or exclusion from the original dictionary. As all sentences in our data were coded manually, we deem the manual evaluation step no longer necessary. Instead, we divided the training corpus into two sets: comments with coded complexity ≤ 2 (not complex) and ≥ 4 (very complex). For each set, we calculated the TF-IDF for all words and selected words with a TF-IDF value larger than a percentile-based threshold. We tried both 95% and 97.5%. In other words, we selected signature words that are characteristic for the comments that are not complex and very complex. With these two sets of signature words, two terms were added to the IC formula:

$$\begin{aligned} IC_{DS} = {} & \mathrm{Sixl} + \mathrm{Discr} + \mathrm{Tent} + \mathrm{Incl} + \mathrm{Cause} + \mathrm{Insig} + \mathrm{Inhib} - -\mathrm{Cert} - -\mathrm{Negate} - -\mathrm{Excl} \\ & + \textit{signature words of the very complex set} - - \text{signature words of the not complex set} \end{aligned} \tag{4}$$

Additivity assumption

The original additivity assumption states that "every instance of every word contributes isomorphically to the output" (Young & Soroka, 2012, p. 209). Many early sentiment dictionaries do assume an isomorphic weighting scheme, e.g., sentiment analysis based on LIWC and General Inquirer. This statement, however, is not technically correct for the IC calculation because not all word categories contribute isomorphically to the output (e.g., words in the category "Sixl" vs. those in "Excl"). It is also not applicable to many off-the-shelf dictionaries that have applied weight to different words, e.g., AFINN or VADER. We therefore restate the additivity assumption in a more general form, so that it reads as follows: "Every instance of every word contributes to the output according to its predefined weight." If the predefined weighting scheme is isomorphic, then our restatement is equivalent to that from Young and Soroka (2012).

We attacked this assumption by modifying the original weighting scheme of the IC score. The original weighting scheme, according to Wyss et al. (2015) was derived to be "in chimes with manual coding" (p. 644) on IC in Suedfeld et al. (1992). We created a modified weighting scheme in chimes with manual coding through machine learning. It is important to note that the input of our machine learning algorithms was still the scores from LIWC, not individual words. By doing so, the machine learning algorithms only adjusted the weight for each LIWC category and kept the other two assumptions intact.

In this analysis, we used a two-level scheme. On the first level, we included exclusively the ten LIWC dimensions put forward by Wyss et al. (2015) to have either a +1 or – 1 weight coefficient (see equation 2). On the second level, we included all available LIWC dimensions, irrespective of whether they had a +1, −1, or 0 weight coefficient in the original weighting scheme (see equation 3).

Four machine learning algorithms were used to address the additivity assumption: linear regression, linear regression with L1 regularization (also known as Lasso regression), the tree-based M5P algorithm, and random forest. For the machine learning approaches on the second level, we decided to remove the word count from the generated features by LIWC. The word count is highly correlated with the resulting IC scores, that is, user comments with higher integrative complexity tend to also be longer. However, as we aim to measure integrative complexity based on the content of the user comments, we removed word count as a feature to not skew the classifiers and their performance. This also helps generalize the trained classifiers, as it removes the overfit to the training data set and thereby increases the usefulness of the trained models when applied to other domains.

Linear regression. As a first step, we used a linear regression model trained to predict the target dimension based on the LIWC categories. Compared to the baseline model, linear regression can set weights in the form of regression coefficients for each of the LIWC categories and therefore increase or decrease the impact of each category on the prediction. The IC formula gets adjusted to:

$$\begin{aligned}\mathrm{IC_{LR}} = {} & \beta_1 * \mathrm{Sixl} + \beta_2 * \mathrm{Discr} + \beta_3 * \mathrm{Tent} + \beta_4 * \mathrm{Inc}l \ + \beta_5 * \mathrm{Cause} + \beta_6 * \mathrm{Insig} + \beta_7 \\ & * \mathrm{Inhib} + \beta_8 * \mathrm{Cert} + \beta_9 * \mathrm{Negate} + \beta_{10} * \mathrm{Excl} + \beta_{11} * \mathrm{Analytic} + \beta_{12} * \\ & \mathrm{Clout} + \ \ldots \end{aligned} \tag{5}$$

The weights are automatically set by training the regression model based on the provided training data while minimizing the ordinary least squares cost function written as:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \tag{6}$$

Lasso regression. The second algorithm tested was linear regression with L1 regularization. Here the cost function gets modified and includes a "Least Absolute Shrinkage and Selection Operator" (Lasso). The cost function can be written as:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \sum_{j=1}^{p}\left|\beta_j\right| \tag{7}$$

By minimizing the extended cost function, features with low weights get eliminated. This adds a feature selection element to the linear regression, which helps to avoid overfitting on the training data.

M5P regression. The third machine learning algorithm we tested was M5P (Quinlan, 1992; Wang & Witten, 1997). It is a tree-based regression algorithm. First, it constructs a decision tree with the goal to minimize the variance of the target variable with each split criterion. It then stops splitting the training set when the variance of the remaining instances in the leaf nodes is below a certain threshold or if the number of instances is below a particular minimum. Finally, a regression model gets trained on the remaining instances which is then used to predict the output. The resulting M5P model is a collection of linear regression models, where the decision tree is used to select the model that is ultimately used for the prediction. One advantage of tree-based algorithms is that they automatically manage variable selection, variable importance, missing values, and normalization (Song & Lu, 2015). Compared to a non-tree based linear regression, it allows for interactions between the features, which is helpful when applied to LIWC categories as many of them are correlated to each other.

Random forest regression. The last algorithm we tested was random forest regression. It adds a random factor to tree-based regression. The random forest model contains *n* decision trees, that each get trained on a random subset of the training data and features. It is therefore an ensemble of tree-based regression models. The final prediction is the average of the value predicted by each individual tree in the ensemble. Adding the random element helps to focus classification on generally important features and therefore, again, avoids overfitting on the training data. Following the usual practice (Lones, 2021), we tuned the three hyperparameters of random forest (number of variables randomly sampled as candidates at each split, split rule, and minimum node size) using the adaptive resampling approach (Kuhn, 2014). We did a sensitivity analysis to study the effect of tuning by comparing the results with that from an untuned random forest regression using the default hyperparameters. Tuning brought about a negligible change in the 2nd decimal place of the accuracy metrics. Nonetheless, we report the results based on the random forest regression with tuned hyperparameters below.

Shotgun approaches

In contrast to our assumption-based approach, a shotgun approach attacks multiple assumptions at the same time. This is currently the default mode of dealing with text analytic tasks. The most important characteristic of this approach is to use full text, instead of scores extracted from LIWC, as the features. Our shotgun approach is an imitation of Van Atteveldt et al. (2021). Their approach consists of a preprocessing step of lemmatizing all texts (using spacy, as stated above) and then a machine learning step.

We used two machine learning algorithms. The first one is the random forest regression as stated above. For this algorithm, lemmatized comments were tokenized and formatted into a document-term matrix for learning. The second one is the convolutional neural network (CNN, a so-called deep learning approach), which Van Atteveldt et al. (2021) demonstrated to have the best agreement with human coders out of all automated content analysis methods. The network architecture we used is the same as Van Atteveldt et al. (2021), which is described as “a relatively standard architecture for document classification” (p. 7). It consists of the following layers:

(1) An embedding layer using the 300-dimensional fastText multilingual embeddings
(2) A convolutional layer that concatenates the embeddings for each 3-word window
(3) A max-pooling layer
(4) A regular densely connected layer that predicts the IC from the pooled features

This network architecture is different from Van Atteveldt et al. (2021) only in that we used the fastText multilingual embeddings, instead of the Amsterdam Embeddings Model, because the latter is not available for English and German.

Table 1. Results of the Cross-Validation for the Different Approaches.

	English		German	
Approach	RMSE	Corr	RMSE	Corr
Baseline (Wyss et al., 2015)	30.30	−.07	32.90	.01
Assumption: Bag-of-words				
Negation (Young & Soroka, 2012)	31.40	−.04	34.40	.02
POS-tagging (Jacobi et al., 2015)	51.30	.11	62.60	.16
POS-tagging (Benamara et al., 2007)	48.60	.15	58.90	.08
Lemmatization (Haselmayer & Jenny, 2017)	22.50	−.06	29.30	.04
Assumption: Domain transferability				
Adj. word choices (Diesner & Evans, 2015) – 5%	25.80	.04	30.10	.10
Adj. word choices (Diesner & Evans, 2015) – 10%	24.60	.06	29.80	.11
Assumption: Additivity				
10 features				
Linear regression [1]	0.93	.30	1.04	.22
Lasso regression	0.93	.30	1.04	.22
M5P	0.75	.64	0.84	.62
Random forest regression	0.72	.68	0.78	.67
All features [2]				
Linear regression [1]	0.83	.52	0.84	.60
Lasso regression	0.83	.52	0.84	.61
M5P	0.76	.62	0.81	.65
Random forest regression	0.70	.70	0.75	.71
Shotgun full-text machine learning				
CNN (fastText Word Embeddings)	0.75	.71	0.84	.69
Random forest	0.76	.73	0.85	.72

Note: RMSE = root mean squared error. POS = part-of-speech. CNN = convolutional neural network.

[1] Please refer to Online Appendix II for the regression coefficients.

[2] The table reports the performance for machine learning models trained without word count as a feature (see section "Additivity assumption").

Evaluation

For the evaluation, we measured the root mean squared error (RMSE) and the correlation of the prediction method with the gold standard (see Table 1 below). The correlation coefficient indicates the general trend between the predicted values and the gold standard. As its magnitude could be influenced by data points with high leverage, e.g., outliers, we also consider the RMSE, which quantifies the average difference between the predicted values and the gold standard. The RMSE is the root of the mean squared error of all predictions. For a test dataset with N datapoints and the dependent variable y it is calculated by equation 8, with the predicted variable y_p and the observed outcome in the gold standard y_{gs}.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_{p,i} - y_{gs,i}\right)^2}{N}} \tag{8}$$

By squaring the residual errors, the RMSE is always positive as well as giving more weight to predictions that have a larger deviation from the actual value in the gold standard. Please note that a lower RMSE indicates a better fit.

Results

Comparing performance

Table 1 shows the RMSE and the correlation of the prediction by each of the methods when compared to the manually coded gold standard. For easier comparison, it reports the results for each of the languages separately.

When compared to the gold standard, the theoretically derived baseline LIWC formula showed no correlation and a high RMSE. This might be due to the fact that this formula was validated against Swiss parliamentary debate contributions, which are quite different from online user comments. This result further emphasizes the need for a thorough validation before applying dictionaries in new contexts.

The first enhancement step, namely the attempt to tackle the bag-of-words assumption by adjusting word choices, only yielded minor improvements over the baseline. While both POS-tagging approaches slightly increased the correlation, removing negated words and lemmatization did not notably improve the correlation. Only lemmatization slightly improved the RMSE (English = 22.5, German = 29.3). While text pre-processing can have a high impact on classification performance, this does not seem to be the case in this particular use case. The results show that text-pre-processing needs to be applied carefully, as it can also have a negative effect.

The second improvement step that dealt with the domain transferability assumption did not increase the performance to acceptable levels either. When looking at the RMSE, the method worked slightly better on the English corpus. However, the correlation is higher for the German model. This is because both performance measures adjust differently depending on the size of the residual error. This shows that the measure that the researcher chooses to optimize performance needs to be carefully selected based on which type of error is more important overall, and that it is often beneficial to look at multiple measures simultaneously. While adjusting the TF-IDF parameter yielded further gains, the overall correlation remained low and RMSE still rather high. Thus, in our case, the low correlation and high error do not seem to be rooted in the lack of domain transferability of the dictionary.

Thirdly, employing machine learning to tackle the additivity assumption was most promising. In the first-level analysis with the ten theoretically derived LIWC features, each of the machine learning approaches improved both the RMSE and the correlation with the prediction variable. While the regression methods (linear and lasso regression) already compared favorably to the LIWC baseline, the tree-based regression methods (M5P and random forest) showed the best performance. Random forest regression had the best RMSE (English = 0.72, German = 0.78) and correlation values (English = 0.68, German = 0.67) of the tested dictionary-plus-supervised-learning approaches. Bringing in features with a zero coefficient in the original weighting scheme (2nd level analysis) brought some additional, albeit not huge, improvement.

To study the weighting procedure in greater detail, we looked at the magnitude of the regression coefficients from the linear regressions (equation 5; see Online Appendix II). The derived weight coefficients are drastically different from those of the original equation (equation 2). For the regression with the 10 LIWC features on the English user comments, for example, when setting the weight coefficient of *Sixl* to – 1, the largest weight coefficient (that of *Excl*) can be as high as +21. This indicates that the original expert-derived weighting scheme by Wyss et al. (2015) with only 0, +1, and −1 is probably too conservative. It suggests that the predefined weights are not realistic, and therefore, attacking the additivity assumption brings about a massive improvement.

Attacking just the additivity assumption had a similar performance as the shotgun full-text machine learning approaches, namely the full-text random forest and the state-of-the-art neural networks approach using word embeddings. The shotgun random forest yielded the best performance of all methods for English (RMSE = 0.76, Correlation = 0.73) and German (RMSE = 0.85, Correlation = 0.72). The neural network approach fell slightly behind in both languages when we measured RMSE (English = 0.75, German = 0.84) and its correlation with the gold standard (English = 0.71, German = 0.69).

Variable importance

The random forest regression model that was combined with the off-the-shelf dictionary additionally allows to generate a variable importance table. This measurement quantifies the reduction in predictive performance of the random forest regression model when a variable of interest is randomly

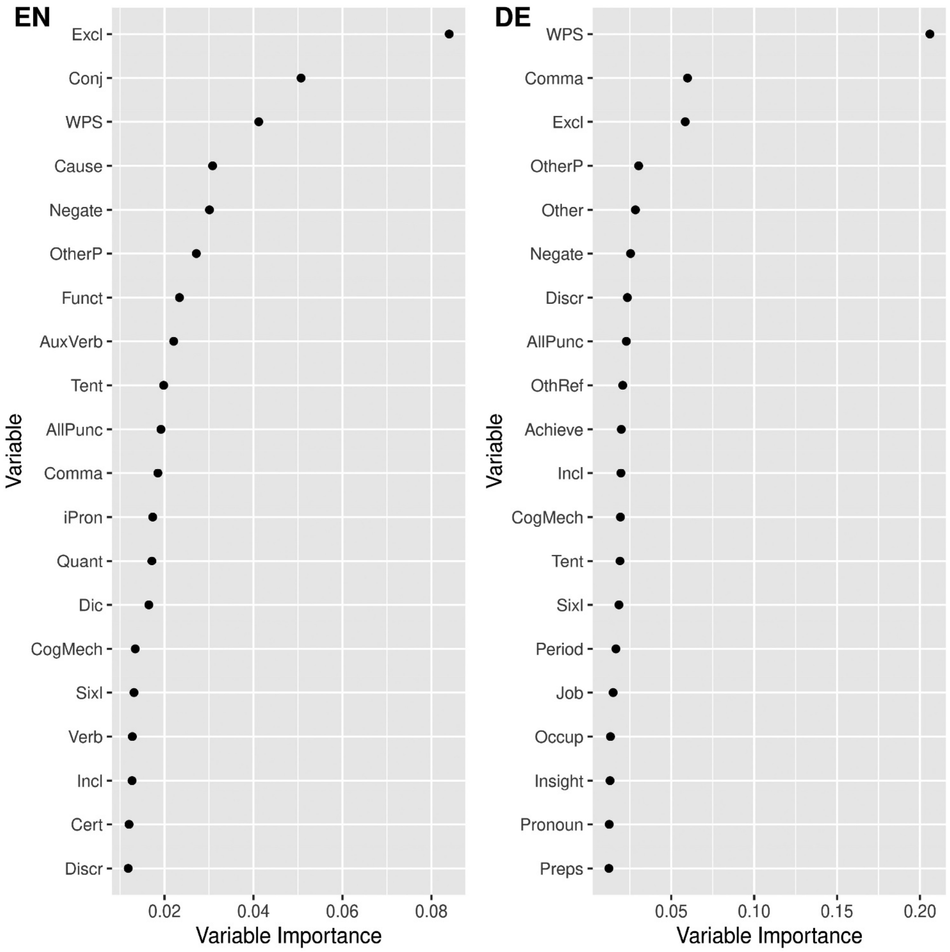

Figure 2. Variable Importance for the English (left) and German (right) Random Forest Models Fed With All LIWC Dictionary Features. *Note*: Achieve = Achievement (*earn, hero, win*), AllPunc = All punctuations, Article (*a, an, the*), AuxVerb = Auxiliary verbs (*am, is, was*), Cause = Causation (*because, effect, hence*), Cert = Certainty (*always, never*), CogMech = Cognitive Processes (*cause, ought*), Comma, Conj = Conjunctions (*and, but, whereas*), Dic = percentage of words in the dictionary, Discr = Discrepancy (*should, would, could)*, Excl = Exclusive (*but, except, without*), Funct = Function words (*no, to, very*), Incl = Inclusive (*with, and, include*), Insight (*think, know*), iPron = Impersonal pronouns (*it, it's, those*), Job (*employ, boss, career*), Negate = Negations (*no, not, never*), Occup = Occupation (work, class, goal), Other = Other grammar words (e.g common verbs), OtherP = Other Punctuation, OthRef = Other References to people (*he, she, they*), Period, Preps = Prepositions (*above, on, to*), Pronoun (*I, we, it*), Quant = Quantifiers (*few, many, much*), Sixl = words > 6 letters, Tent = Tentative (*maybe, perhaps*), Verb = Common verb (*ask, eat, cry*), WPS = Words per sentence.

shuffled along rows. Applying the original terminology from Breiman's landmark paper (Breiman, 2001), the variable of interest is "noised" (being turned into meaningless random noise). A greater reduction in performance after a variable being "noised" indicates that the variable is relatively more important for generating the prediction. It is important to note that the value is comparative only across the variables in the random forest model and thus there is no cutoff point. Due to the complex interactions among variables in a random forest regression model, the metric does not tell us anything about the relative importance of variables in the case of a simple linear combination of variables (e.g., equation 2). Nonetheless, this makes it possible to find those LIWC categories that were used (most heavily) by the regression trees in the random forest ensemble to generate the prediction. Figure 2 shows the variable importance for the random forest classifiers fed with all available LIWC dimensions (2nd level analysis), trained on the German and English corpus.

In both models the top categories comprise several ones that also appear in the baseline LIWC formula used by Wyss et al. (2015), including for example, “Excl,” “Discr,” “Sixl,” “Incl,” “Negate,” “Tent,” and “Cause.” This suggests that the resulting classification model is closely linked to the existing theory, which further validates this approach. The conjunctions (“Conj”) that are rather important in the English model have also previously been part of theoretically derived LIWC-based operationalizations of integrative complexity (Abe, 2011). Moreover, formal features such as punctuation (e.g., “AllPunc,” “OtherP,” “comma”) or words per sentence (“WPS”) seem to carry important information about the integrative complexity of the user comments. Additional top categories such as job-related words, quantifiers, or achievement words have a further impact on explaining the operationalization of integrative complexity in our case study.

The variable importance analysis presented in Figure 2 is tied closely to our use case and especially our dataset. Its main use is to open the black box to diagnose the random forest model. This allows for further insights into our dataset and the specific operationalization of integrative complexity that we have based our study on. Yet, further research is required to show which categories carry importance when this method is transferred to other datasets. This reiterates our point for the thorough validation of instruments when carrying them over to new datasets or application domains.

Conclusion

Echoing previous research (Van Atteveldt et al., 2021), our study reemphasizes that off-the-shelf dictionaries such as LIWC cannot be applied to different research questions and contexts without (re) validation against a manually coded gold standard. In our case, we aimed to transfer a theoretically derived and empirically validated off-the-shelf dictionary measurement for the integrative complexity of parliamentary debate contributions (Wyss et al., 2015) to the classification of the same phenomenon in a diverse set of online user comments from different communication arenas and countries. The analysis showed that if the (re)validation step is ignored, we would apply a measurement tool with almost no correlation to the (manually coded) construct that we want to measure. This reminds us that off-the-shelf dictionaries are theory-informed methods developed for very specific situations and restates the important gate-keeping function of manual validation that has been stressed by social scientists already early on (Grimmer & Stewart, 2013).

However, the current literature does not provide a compelling answer to what should be done next when the gate is closed for an off-the-shelf dictionary-based method. The suggested default alternative is state-of-the-art full-text machine learning (González-Bailón & Paltoglou, 2015; Van Atteveldt et al., 2021). Yet, while this approach has been shown to work well in practice, it addresses several weaknesses of off-the-shelf dictionaries at the same time, namely the bag-of-words, domain transferability, and additivity assumption, without explaining which of these problems is tackled by the machine learning. This shotgun approach thus tells us very little about why the off-the-shelf dictionary fails in the first place.

The first innovation of this paper is to propose and follow a general assumption-based error analysis framework that tackles the three assumptions one at a time to pinpoint the major source of weakness in dictionary-based methods. We demonstrate that our framework is equally applicable to both English and German. For our case of measuring integrative complexity in online user comments, we found that compared to the LIWC baseline, the largest improvement in performance comes from attacking the additivity assumption. This is particularly interesting because this assumption has to date received comparatively little attention in the literature (Young & Soroka, 2012) as opposed to the bag-of-words (Chan et al., 2021) and domain transferability assumptions (González-Bailón & Paltoglou, 2015) that many researchers are deeply concerned with. By just attacking the additivity assumption, we obtained an improvement in accuracy that is on par with that of the shotgun full-text machine learning approach (Van Atteveldt et al., 2021). While this promising performance may be specific to the integrative complexity of online user comments, the general conclusion regarding the additivity assumption is likely transferable to other use cases. An additional analysis using the off-the-shelf sentiment dictionary AFINN (Nielsen, 2011) and a large corpus of annotated tweets (see Online Appendix III) demonstrates the generalizability of our framework as well as our findings. Although the task is quite different, e.g., in measuring sentiment vs. integrative complexity and in that the sentiment analysis is a classification rather than a regression task, our framework can be applied in this context without any problems. Again, attacking the additivity assumption brings the greatest improvement in dictionary performance compared to attacking the bag-of-words and domain transferability assumption, respectively. However, while this mirrors the overall conclusion of our main analysis in this paper, one difference is that the improvement is comparatively more modest and the performance of the dictionary-plus-supervised-learning approach falls short of the power of shotgun full-text machine learning approaches more clearly in the sentiment case. The former can be explained by the fact that AFINN has a reasonable performance in measuring the sentiment of the analyzed tweets in the first place, leaving less room for improvement. This may be due to the fact that sentiment in tweets is, after all, easier to measure automatically than a more complex content-related construct such as integrative complexity – or because the AFINN dictionary is better tailored to measuring sentiment than LIWC is to capturing integrative complexity. In particular, the AFINN dictionary implements a more sophisticated weighting scheme, which has weight coefficients spanning from – 5 to +5, that may already address a good portion of the additivity assumption in this case. Nonetheless, our systematic error analysis suggests that the additivity assumption is rather unrealistic in many circumstances. Even when dictionary-based methods assign different weights (e.g., AFINN), those weights might work differently on new data or in different contexts (Young & Soroka, 2012). Thus, in conducting dictionary-based content and sentiment analyses alike, social scientists should give intensive thought to how different weights could be applied to different indicators in order to improve the classification of the data.

The second innovation of this paper is also informed by our error analysis framework. The comparison of our dictionary-plus-supervised-learning method to tackle the additivity assumption to the state-of-the-art shotgun full-text machine learning shows that different "approaches to automated content analysis need not compete" (Dun et al., 2020, p. 14) since they can generate very satisfactory performances when used in combination. Bundling (theory-informed) off-the-shelf dictionaries with machine learning algorithms that attack the additivity assumption in these dictionaries can generate valid and interpretable models for automated text classification tasks. Compared to full-text machine learning, the advantage of this combinatory approach lies primarily in the interpretability of those models (Lipton, 2018). In contrast to the shotgun approach, a classifier that combines off-the-shelf dictionaries with machine learning provides substantial information about which dictionary categories are important in predicting the variable – in a way that is understandable by human beings (Zhang et al., 2020). This simpler approach thus supplies additional insights into the investigated variable which can be used for theory development in the field, that is, it "can help us to revise old or build new theory given empirical results from a machine learning model" (Radford & Joseph, 2020, p. 8). In so doing, the interpretability of the dictionary-plus-supervised-learning approach also simplifies the refinement of subsequent theoretically driven operationalizations. In our case, the random forest classifier that was combined with the LIWC categories delivered both confirmatory and novel insights into which groups of words exactly make an online argument more complex (Figure 2).

Combining word lists with machine learning may thus be one way to meet current calls for a more strongly theory-driven automated analysis of large text corpora (Baden et al., 2020). In line with Rai (2019), we propose to term this "glass box" machine learning as opposed to the usual "black box" machine learning. In some cases, glass box models might not have the better accuracy compared to their black box counterparts, i.e., the full-text machine or deep learning approaches. However, trading the (small) penalty in accuracy for explainable insights that pay off in theory development may be a reasonable price to pay for many social scientists. In our case, this penalty was even negligible, and on the German user comments, the CNN approach actually performed worse than the random forest glass box model trained with all the LIWC features.

Yet, combining an off-the-shelf dictionary with supervised machine learning does not come without disadvantages. Those include, for example, the need for a training data set of adequate size, which can be time-intensive and costly to produce. As numerous studies (e.g., Chan et al., 2021; Grimmer & Stewart, 2013; Van Atteveldt et al., 2021) have shown the importance of properly validating automated text analysis tools, though, training human coders and manually scoring a gold standard set for such validation is always necessary. Therefore, producing a training data set for a dictionary-plus-supervised-learning approach does not require as much extra effort as it might seem at first, especially since pre-processing the data with an off-the-shelf dictionary can reduce the quantity of hand-coded data that would be needed for full-text machine learning.

Furthermore, as it relies on off-the-shelf dictionaries, the proposed dictionary-plus-supervised-learning approach may only be applicable to a limited number of languages. While there are many dictionaries available for common languages like English and German, this choice is rather restricted when it comes to languages that are used less widely, such as Dutch (Boukes et al., 2020). In this context, the present study is constrained by the fact that it tested different automated text classification tools only on user comments in two Germanic languages, while disregarding others (Baden et al., 2021). Thus, the analysis cannot speak to whether these methods and the derived conclusions are equally applicable to languages from other families, such as Sino-Tibetan or Semitic languages. In particular, applying the proposed dictionary-plus-supervised-learning approach to data in these languages could be challenging because dictionaries are scarcely available in these tongues. Our analysis on the German user comments has shown that translated dictionaries can also lead to acceptable performance levels, at least for related languages like English and German, even though this requires additional manual work on the part of the researcher. Dictionary translation alone has been shown to be viable for comparative automated content analysis of content in European languages (Lind et al., 2019). However, it remains an open question whether this dictionary translation approach would also work for a) the added machine learning procedure, and 2) languages that are more distant from each other and belong to different families. Thus, future research should focus more strongly on testing and developing automated classification tools for non-Germanic as well as less commonly spoken languages (Baden et al., 2021; Boukes et al., 2020), which would also help to further cross-national computational social science research (Maier et al., 2021).

Similarly, even though the previously reported robustness check (Online Appendix III) has shown that our framework and conclusions are equally applicable when analyzing sentiment rather than integrative complexity, the results of this study need to be further substantiated by examining different constructs in different topical contexts.

Finally, we acknowledge that the skills needed to employ machine learning as a social scientist might pose a challenge. Therefore, we have provided guidelines with our framework and have also shared our code[2] as a template for researchers who would like to adapt any of the approaches. However, going forward, those methods should also be made more accessible by providing user-friendly software that can apply machine learning solutions similar to existing text analysis software such as LIWC. Moreover, as the requirements for future researchers are developing further in the direction of automated text analysis, training them in the tools and skills needed in this area will become increasingly important.

In advocating for the combination of interpretable machine learning algorithms with off-the-shelf dictionaries, we align ourselves with researchers who have stressed "that the rapid development of computational methods has not been accompanied by an equally strong emphasis on theoretical developments within the scholarly community" (Waldherr et al., 2021, p. 3, see also Van Atteveldt et al., 2019). Ultimately, most computational text analysis in the social sciences is concerned with a more specific theoretical problem rather than the classification task itself. In line with other

[2] See online appendix hosted on https://doi.org/10.17605/OSF.IO/578MG

colleagues (Grimmer et al., 2021; Radford & Joseph, 2020), we therefore advocate the use of machine learning to advance theory building in social science. Our use case has shown that selecting an interpretable machine learning algorithm can be more valuable for the researcher than just selecting the method "that optimizes performance for their particular research task" (Grimmer et al., 2021, p. 404). Knowing why our machine learning algorithm has made certain predictions and how the model's measurement categories relate to the theoretical concepts that we want to capture helps us learn more about the problem that we are interested in in the first place, and thus contributes to theory development in the field.

Acknowledgments

We want to thank André Bächtiger, Simon Beste, and Dominik Wyss for sharing information about their previous studies.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Deutsche Forschungsgemeinschaft [260291564].

ORCID

Timo Dobbrick http://orcid.org/0000-0002-6252-1157
Chung-Hong Chan http://orcid.org/0000-0002-6232-7530
Hartmut Wessler http://orcid.org/0000-0003-4216-5471

References

Abe, J. A. A. (2011). Changes in Alan Greenspan's language use across the economic cycle: A text analysis of his testimonies and speeches. *Journal of Language and Social Psychology*, *30*(2), 212–223. https://doi.org/10.1177/0261927X10397152

Baden, C., Kligler-Vilenchik, N., & Yarchi, M. (2020). Hybrid content analysis: Toward a strategy for the theory-driven, computer-assisted classification of large text corpora. *Communication Methods and Measures*, *14*(3), 165–183. https://doi.org/10.1080/19312458.2020.1803247

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2021, May 27–31). Three gaps in computational methods for social sciences: A research agenda. [Conference presentation]. *71th Annual International Communication Association Conference* virtual.

Baker-Brown, G., Ballard, E. J., Bluck, S., de Vries, B., Suedfeld, P., & Tetlock, P. E. (1992). The conceptual/integrative complexity scoring manual. In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic content analysis* (pp. 401–418). Cambridge University Press.

Benamara, F., Cesarano, C., Picariello, A., Recupero, D. R., & Subrahmanian, V. S. (2007). Sentiment analysis: Adjectives and adverbs are better than adjectives alone. , *7*, 203–206.

Beste, S., & Wyss, D. (2014, September 3–6). Cognitive complexity as a proxy for high quality deliberation? A theoretical and empirical exploration of cognitive complexity and deliberative quality in the EuroPolis discussions [Paper presentation]. *European Consortium for Political Research General Conference*, Glasgow, United Kingdom. Colchester, United Kingdom: ECPR Press.

Boukes, M., van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What's the tone? Easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures*, *14*(2), 83–104. https://doi.org/10.1080/19312458.2019.1671966

Boumans, J. W., & Trilling, D. (2015). Taking stock of the toolkit. *Digital Journalism*, *4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chan, C. H., Bajjalieh, J., Auvil, L., Wessler, H., Althaus, S., Welbers, K., van Atteveldt, W., & Jungblut, M. (2021). Four best practices for measuring news sentiment using 'off-the-shelf' dictionaries: A large-scale p-hacking experiment. *Computational Communication Research*, *3*(1), 1–27. https://doi.org/10.5117/CCR2021.1.001.CHAN

Diesner, J., & Evans, C. S. (2015). Little bad concerns: Using sentiment analysis to assess structural balance in communication networks. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 -ASONAM '15* Paris, France. New York, NY, United States: Association for Computing Machinery, 342–348. https://doi.org/10.1145/2808797.2809403

Dun, L., Soroka, S., & Wlezien, C. (2020). Dictionaries, supervised Learning, and media coverage of public policy. *Political Communication*, *38*(1–2), 140–158. https://doi.org/10.1080/10584609.2020.1763529

González-Bailón, S., & Paltoglou, G. (2015). Signals of public opinion in online communication. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 95–107. https://doi.org/10.1177/0002716215569192

Grimmer, J., Roberts, M., & Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, *24*(1), 395–419. https://doi.org/10.1146/annurev-polisci-053119-015921

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

Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & Quantity*, *51*(6), 2623–2646. https://doi.org/10.1007/s11135-016-0412-4

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo. https://doi.org/10.5281/zenodo.1212303

Jacobi, C., van Atteveldt, W., & Welbers, K. (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89–106. https://doi.org/10.1080/21670811.2015.1093271

Kenneth, B., & Matsuo, A. (2020). *Spacyr: An R wrapper for spaCy*. Retrieved May 21st, 2021, from https://spacyr.quanteda.io

Kesting, N., Reiberg, A., & Hocks, P. (2018). Discourse quality in times of populism: An analysis of German parliamentary debates on immigration policy. *Communication & Society*, *31*(3), 77–91. https://doi.org/10.15581/003.31.3.77-91

Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel*. Institute for Simulation and Training, University of Central Florida. https://stars.library.ucf.edu/istlibrary/56

Kuhn, M. (2014). *Futility analysis in the cross-validation of machine learning models*. ArXiv. https://arxiv.org/abs/1405.6974

Lind, F., Eberl, J. M., Heidenreich, T., & Boomgaarden, H. G. (2019). When the journey is as important as the goal: A roadmap to multilingual dictionary construction. *International Journal of Communication*, *13*, 4000–4020. https://ijoc.org/index.php/ijoc/article/view/10578

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 1–27. https://doi.org/10.1145/3236386.3241340

Lones, M. A. (2021). *How to avoid machine learning pitfalls: A guide for academic researchers*. ArXiv. https://arxiv.org/abs/2108.02497

Maier, D., Baden, C., Stoltenberg, D., De Vries-Kedem, M., & Waldherr, A. (2021). Machine translation vs. multilingual dictionaries: Assessing two strategies for the topic modeling of multilingual text collections. *Communication Methods and Measures*. https://doi.org/10.1080/19312458.2021.1955845.

Moore, A., Fredheim, R., Wyss, D., & Beste, S. (2020). Deliberation and identity rules: The effect of anonymity, pseudonyms and real-name requirements on the cognitive complexity of online news comments. *Political Studies* *69*(1), 45–65 . https://doi.org/10.1177/0032321719891385

Muddiman, A., McGregor, S. C., & Stroud, N. J. (2019). Re)Claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication*, *36*(2), 214–226. https://doi.org/10.1080/10584609.2018.1517843

Ng, A. (2018). *Machine learning yearning: Technical Strategy for AI Engineers, In the Era of Deep Learning* (deeplearning.ai).

Nielsen, F. Å. (2011). *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs*. ArXiv. https://arxiv.org/abs/1103.2903

Owens, R. J., & Wedeking, J. P. (2011). Justices and legal clarity: Analyzing the complexity of U.S. Supreme Court opinions. *Law & Society Review*, *45*(4), 1027–1061. https://doi.org/10.1111/j.1540-5893.2011.00464.x

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *LIWC2007: Linguistic inquiry and word count*. Austin, Texas: liwc.net.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of liwc2015*. University of Texas at Austin. https://repositories.lib.utexas.edu/handle/2152/31333

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Quinlan, J. R. (1992). Learning with continuous classes. *Proceedings of Australian Joint Conference on Artificial Intelligence*, *92*, 343–348 http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.34.885.

Radford, J., & Joseph, K. (2020). Theory in, theory out: The uses of social theory in machine learning for social science. *Frontiers in Big Data*, *3*(18), 18. https://doi.org/10.3389/fdata.2020.00018

Rai, A. (2019). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F. (2016). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, *5*(1), 23. https://doi.org/10.1140/epjds/s13688-016-0085-1

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2–3), 140–157. https://doi.org/10.1080/19312458.2018.1455817

Song, -Y.-Y., & Lu, Y. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.1002-0829.215044

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting incivility and impoliteness in online discussions: Classification approaches for German user comments. *Computational Communication Research*, *2*(1), 109–134. https://doi.org/10.5117/CCR2020.1.005.KATH

Suedfeld, P., Tetlock, P. E., & Streufert, S. (1992). Conceptual/integrative complexity. In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic content analysis* (pp. 393–400). Cambridge University Press.

van Atteveldt, W., Margolin, D., Shen, C., Trilling, D., & Weber, R. (2019). A roadmap for Computational Communication Research. *Computational Communication Research*, *1*(1), 1–11. https://doi.org/10.5117/CCR2019.1.001.VANA

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

Waldherr, A., Geise, S., Mahrt, M., Katzenbach, C., & Nuernbergk, C. (2021). Toward a stronger theoretical grounding of computational communication science: How macro frameworks shape our research agendas. *Computational Communication Research*, *3*(2), 1–28. https://doi.org/10.5117/CCR2021.02.002.WALD

Wang, Y., & Witten, I. H. (1997). Inducing model trees for continuous classes. *Proceedings of the Ninth European Conference on Machine Learning 9* (1) , 128–137. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.9768.

Wyss, D., Beste, S., & Bächtiger, A. (2015). A decline in the quality of debate? The evolution of cognitive complexity in Swiss parliamentary debates on immigration (1968–2014). *Swiss Political Science Review*, *21*(4), 636–653. https://doi.org/10.1111/spsr.12179

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, *29*(2), 205–231. https://doi.org/10.1080/10584609.2012.671234

Zhang, Y., Tiňo, P., Leonardis, A., & Tang, K. (2020). *A Survey on Neural Network Interpretability. arXiv preprint* arXiv:2012.14261. https://arxiv.org/abs/2012.14261