Lind, Fabienne; Heidenreich, Tobias; Kralj, Christoph; Boomgaarden, Hajo G.

Greasing the wheels for comparative communication research: Supervised text classification for multilingual corpora

Computational Communication Research

Provided in Cooperation with:
WZB Berlin Social Science Center

Suggested Citation: Lind, Fabienne; Heidenreich, Tobias; Kralj, Christoph; Boomgaarden, Hajo G. (2021) : Greasing the wheels for comparative communication research: Supervised text classification for multilingual corpora, Computational Communication Research, ISSN 2665-9085, Amsterdam University Press, Amsterdam, Vol. 3, Iss. 3, pp. 1-30, https://doi.org/10.5117/CCR2021.3.001.LIND

Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 3.3 (2021) 1-30
HTTPS://DOI.ORG/10.5117/CCR2021.3.001.LIND

ARTICLE

Greasing the wheels for comparative communication research: Supervised text classification for multilingual corpora

Fabienne Lind
Department of Communication, University of Vienna

Tobias Heidenreich
Department of Communication, University of Vienna

Christoph Kralj
Department of Computer Science, University of Vienna

Hajo G. Boomgaarden
Department of Communication, University of Vienna

Abstract

Employing supervised machine learning for text classification is already a resource-intensive endeavor in a monolingual setting. However, facing the challenge to classify a multilingual corpus, the cost of producing the required annotated documents quickly exceeds even generous time and financial constraints. We show how tools like automated annotation and machine translation can not only efficiently but also effectively be employed for the classification of a multilingual corpus with supervised machine learning. Our findings demonstrate that good results can already be achieved with the machine translation of about 250 to 350 documents per category class and language and a dictionary in just one language, which we perceive as a realistic scenario for many projects. The methodological strategy is applied to study migration frames in seven languages (news discourse in seven European countries) and discussed and evaluated for its usability in comparative communication research.

Keywords: multilingual content analysis, text classification, comparative communication research, supervised machine learning, machine translation

The goal of deductive types of automated content analysis is the classification of documents (i.e., texts) into classes of predefined categories to infer meaning from those documents. Employing supervised machine learning (SML) for such classification has gained popularity in many areas of the social sciences (Boumans & Trilling, 2016; Grimmer & Stewart, 2013). In communication science, several methodological articles provide guidance for the sound implementation of SML (e.g., Burscher et al., 2014; Pilny et al., 2019; Scharkow, 2013). While acknowledging their fundamental contributions, it is evident that their focus lies on monolingual corpora (i.e., document collections). Hence, while the discipline has repeatedly emphasized the necessity for more country-comparative research (e.g., Boomgaarden & Song 2019; Esser & Hanitzsch, 2012), so far few transfer SML state-of-the-art methods to research dealing with documents in multiple languages (but see: Courtney et al., 2020). Among the projects that do allow comparative communication research are, first of all, large-scale survey programs (e.g., EES, CSES, ISSP, ESS). The few projects that afforded to include a content analysis on a larger comparative scale typically relied on resource-intensive human coding to classify documents for the countries under study (e.g., EES, Banducci et al., 2014; NEPOCS, Hopmann et al., 2016). It is precisely such projects that would benefit if automated content analysis methods were made fitter for comparative communication research (e.g., Baden et al., 2020). Replacing human coding efforts in parts with automated coding is at best the necessary kickstart for small initiatives that cannot afford to have the complete (multilingual) material coded by human coders (Grimmer & Stewart, 2013, p. 268).

While it is true that automated classification methods such as SML can save on the cost of manually coding the full-text material, it would be wrong to assume that SML can solve all resource issues. There is, first of all, the challenge to collect sufficient amounts of high-quality annotated documents to train and test a text classifier, a process that typically requires human coding or designing an alternative automated coding instrument such as a dictionary for the annotation of documents, both of which are usually resource-intensive undertakings (e.g., Young & Soroka, 2012). In the multilingual case — a setting researchers will often face in a comparative study — the text material that is classified includes different languages.

Employing SML here can be approached by providing a large number of annotated documents in each language. The task of creating such high-quality learning material, however, is much more demanding than it would be in a monolingual setting. In addition to the need to have sufficient annotated documents available for each language, there is the challenge to pay attention to sample equivalence, construct equivalence, measurement equivalence, and procedural equivalence (Esser & Vliegenthart, 2017; Rössler, 2012) when selecting and annotating the documents and when using them for comparative communication research purposes.

Alternatively, to circumvent some of the mentioned resource-related problems, first studies in the fields of political communication (Courtney et al., 2020; Loftis & Mortensen, 2020) and computational linguistics (e.g., Balahur & Turchi, 2014; Banea et al., 2008) tested strategies that rely heavily on machine translation and/or annotated documents for one language. In complementing this type of work, the goal of this contribution is to *evaluate an approach to train high-quality SML classifiers for comparative research that keeps the negative impact of biased training documents and translation errors as well as the costs for annotation and machine translation to a minimum*. To that end, we propose a strategy, the utility of which we outline and demonstrate in this manuscript. In short, in this new strategy, not the full corpus but only parts (document samples for each language) are machine-translated. Validated English-language dictionaries are then applied to annotate the machine-translated samples. The annotation decision is projected to the original language version of the documents, and the related multilingual documents subsequently serve to train and test classifiers for different languages. The strategy to train classifiers with dictionary annotated data is further tested with a separate human-annotated data set. In essence, the strategy provides one answer to the question of how tools like automated annotation and machine translation can be most effectively (i.e., high-quality) but also most efficiently (i.e., resource-saving) employed for the classification of a multilingual corpus with SML. When outlining our and other strategies to tackle this question, we keep the above-mentioned equivalence requirements of comparative communication research in mind and discuss them from this perspective. We try to apply an open science approach where possible. Scripts for dictionary annotation and for classifier training and testing as well as results are published in the online repository for this paper.[2]

To showcase our strategy, we work with news media articles (N = 138,388) dealing with migration in seven languages published in Germany, Hungary, Poland, Romania, Spain, Sweden, and the UK between

January 2017 and November 2018. The four categories that we classify with SML are migration-related frames, the economy & budget, labor market, the security, and welfare frame, all of which are widely studied concepts in the media and migration literature (e.g., Caviedes, 2015; Strömbäck et al., 2017). We chose migration not least due to its global and cross-border character, which makes it an especially relevant field for comparative communication research (Eberl et al., 2018). By contributing to the advancement of automated content analysis procedures for comparative communication research, this manuscript facilitates research on media discourses about migration, but most importantly also research on numerous other topics studied from a comparative perspective, such as climate change (Reber, 2019), EU elections (Schuck et al., 2013), or populism (Gründl, 2020).

Automated Content Analysis for Comparative Research

Comparative research has been presented as a troublesome methodological challenge, but one that is vital for communication research (Esser & Hanitsch, 2012; Livingstone, 2003; Rössler, 2012). The comparative perspective is useful if not necessary to study the transnational dimension of phenomena related to communication processes, to test theories' applicability beyond the individual case (i.e., system, culture, markets, country), to assess theories' contextual boundary conditions, to learn about and explain similarities and differences of cases, and to raise awareness for the contexts of other cases (Esser & Hanitsch, 2012; Livingstone, 2003). Automated content analysis methods can contribute to such research goals when they effectively assist researchers to study large text corpora from different cases that are difficult or almost impossible to handle with human annotations (Grimmer & Stewart, 2013, p. 268).

Of course, the ability of automated tools to reliably process large amounts of data alone is not the end of all problems. In a comparative setting, many methodological questions revolve around dealing with the multilingualism of the text data.[3] Several recent studies have thus made contributions to the advancement of automated content analysis specifically for comparative research in communication science. Approaches suitable for cross-lingual topic extraction (Chan et al., 2020), polylingual topic modeling (Lind et al., 2021), and the scaling of documents in different languages (Watanabe, 2020) were recently presented. For deductive top-down text classification tasks, which are the focus of this manuscript, not only SML but also fully

rule-based approaches like dictionary methods are typically considered (Grimmer & Stewart, 2013). In such dictionaries, it is solely carefully selected keywords that are responsible for capturing all textual patterns that point to the searched categories in each language. When constructed and used for the search in multilingual corpora, the "functional equivalence" (Cohen, 2012, p. 540) of the keywords across languages is thus crucial for the comparability of the results across cases. The few efforts that have been made in this direction in recent years (Baden & Stalpouskaya, 2015; Lind et al., 2021; Proksch et al., 2019) showed that keyword selection can be partly supported automatically, but that the construction of a multilingual dictionary remains overall a manual labor-intensive endeavor. Some studies therefore machine-translated the multilingual documents into one language first and classified the full corpus with a dictionary in that language (e.g., Boot, 2021).[4] If researchers do not want to define all classification rules (e.g., keywords) in advance (Baden et al., 2020) or like to benefit from the often-demonstrated performance lead of SML over dictionary methods (van Atteveldt et al., 2021), an SML approach is useful.

Text Classification with Supervised Machine Learning

When a classification task is approached with SML, human (and dictionary) classification efforts are augmented by classification algorithms. These algorithms learn the correct assignment of classes from so-called annotated data sets, which include examples for each class. If SML is used for the classification of text documents, the first step is the collection of annotated documents, which involves the assignment of classes to documents either manually by human coders or automatically, for example, by dictionaries. The documents are then transferred into a numerical data format to make them accessible for computational analysis (Grimmer & Stewart, 2013). Possible features include word counts, TF-IDF scores, topic probability scores derived from topic modeling, etc. (Pilny et al., 2019). To facilitate learning, feature selection techniques can be used that aim at the exclusion of less informative features (Deng et al., 2019). In the next step, algorithms learn from relationships between the selected features and classes of so-called "training documents," relating to one part of the annotated documents. To evaluate the classification performance of a trained classifier, the classifier predicts the classes for so-called "test documents" (also annotated documents but not part of the training documents). This prediction is based only on the document features of the test documents; available annotations

are invisible to the classifier. The classifier predictions are then compared to the annotations of the test documents. The best possible prediction is when the classifier succeeds in replicating the annotations of the test documents.[5] Such training and test documents are also used to optimize hyperparameters and to assess the performance of trained classifiers. Once a certain performance is reached, the classifier can be used to annotate other, previously unseen documents.

Communication research has made progress in offering recommendations for the thorough use of SML for text classification. Pilny et al. (2019) provide a detailed step-by-step SML overview with a focus on reliability and validity testing. Others presented methods to deal with multiclass classification (Loftis & Mortensen, 2020; Sebők & Kacsuk, 2020), to use word embeddings rather than a bag-of-words approach (Rudkowski et al., 2018), or to work with deep learning algorithms (van Atteveldt et al., 2021). While providing for important refinements to SML-based content analysis approaches in communication science, these contributions focus solely on single-language corpora.

Supervised Classification for Multilingual Corpora

For SML classification tasks, as described above, technically speaking, it does not matter which language the numerical representations of texts originally come from. The SML techniques for text classification have basically the flexibility to be applied to any language. This is also reflected in the linguistic diversity of recently published studies. They teach algorithms topic classification for Arabic (Alkhair et al., 2019), Chinese (Chang & Masterson, 2019), Croatian (Karan et al., 2016), Hungarian (Sebők & Kacsuk, 2020), or Danish texts (Loftis & Mortensen, 2020). The shortage of cross-country comparative studies, therefore, is not due to the capabilities of the algorithms. Rather, the bottleneck seems to be the availability of suitable training and test data sets. Cross-national research programs like the Comparative Agendas Project (Baumgartner et al., 2019) or the Manifesto Project (Volkens et al., 2015) are useful sources for such annotated documents. Such data treasures have, of course, their natural limits. If a research question cannot be answered based on available annotated corpora, researchers must create their own data sets. Multiple strategies can be considered (see Figure 1 for a visualization).

Strategy A: Annotated Documents in All Languages, No Translation, Training of One Classifier per Language

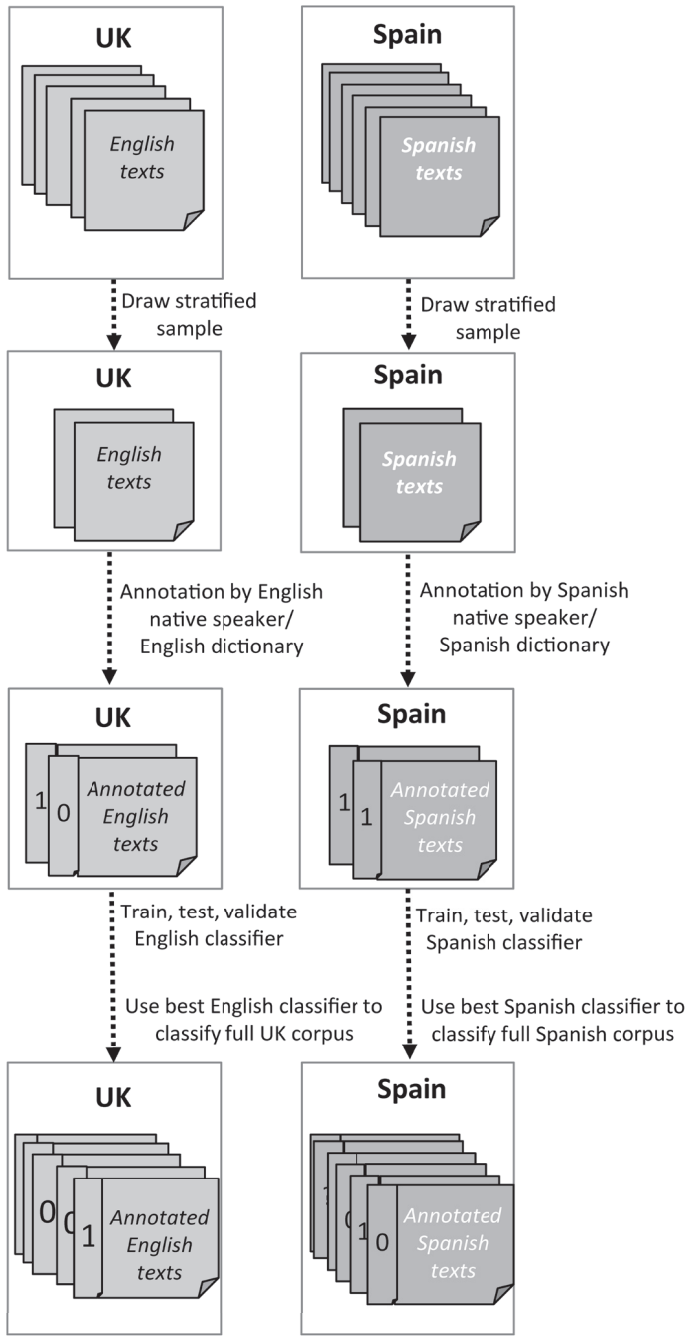
The first strategy is to annotate parts of the full multilingual corpus (one sample of documents per language) in their respective original language. Native speakers are hired to code parts of the material in their respective native language, or multilingual dictionaries are available (or can be constructed) to annotate the documents automatically per language. The annotated multilingual documents are used to train and test classifiers, one per language. In principle, manual coding for several languages in parallel is the procedure of projects such as the above-mentioned Comparative Agendas Project (Baumgartner et al., 2019), the Manifesto Project (Volkens et al., 2015), or the European Election Study (Banducci et al., 2014).

Strategy B: Annotated Documents in One Language, One Classifier for Translated Documents

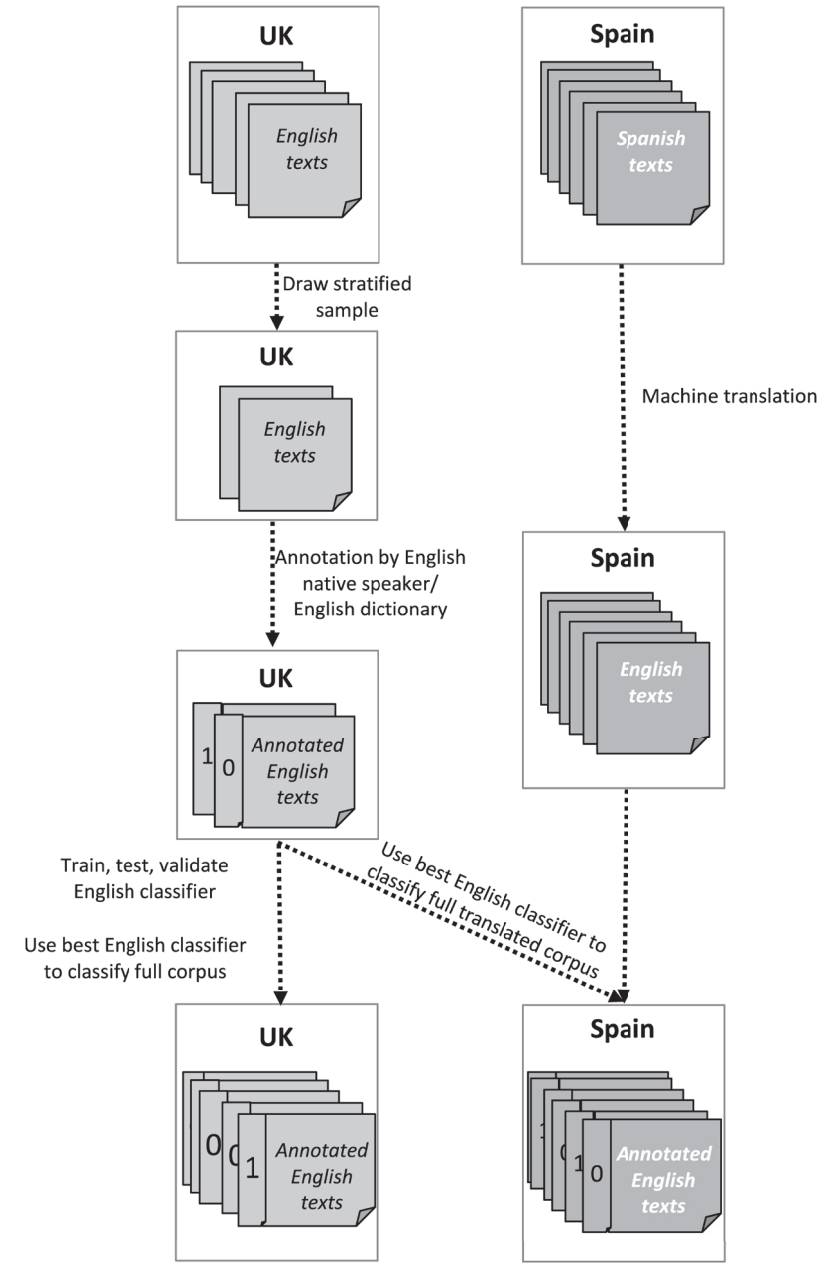
Another strategy to classify a multilingual corpus is to annotate only documents for those languages where annotation means are available (e.g., only a sample of the English documents), to annotate these documents manually or automatically, and to use them to train and test a classifier. To apply this classifier for the scoring of the full multilingual corpus, the non-English documents are first (machine) translated into English. This strategy was tested by Loftis and Mortensen (2020, Supplemental Material, Appendix D).

Strategy C: Annotated Documents in One Language, Translation of Annotated Documents into Several Languages, Training of One Classifier per Language

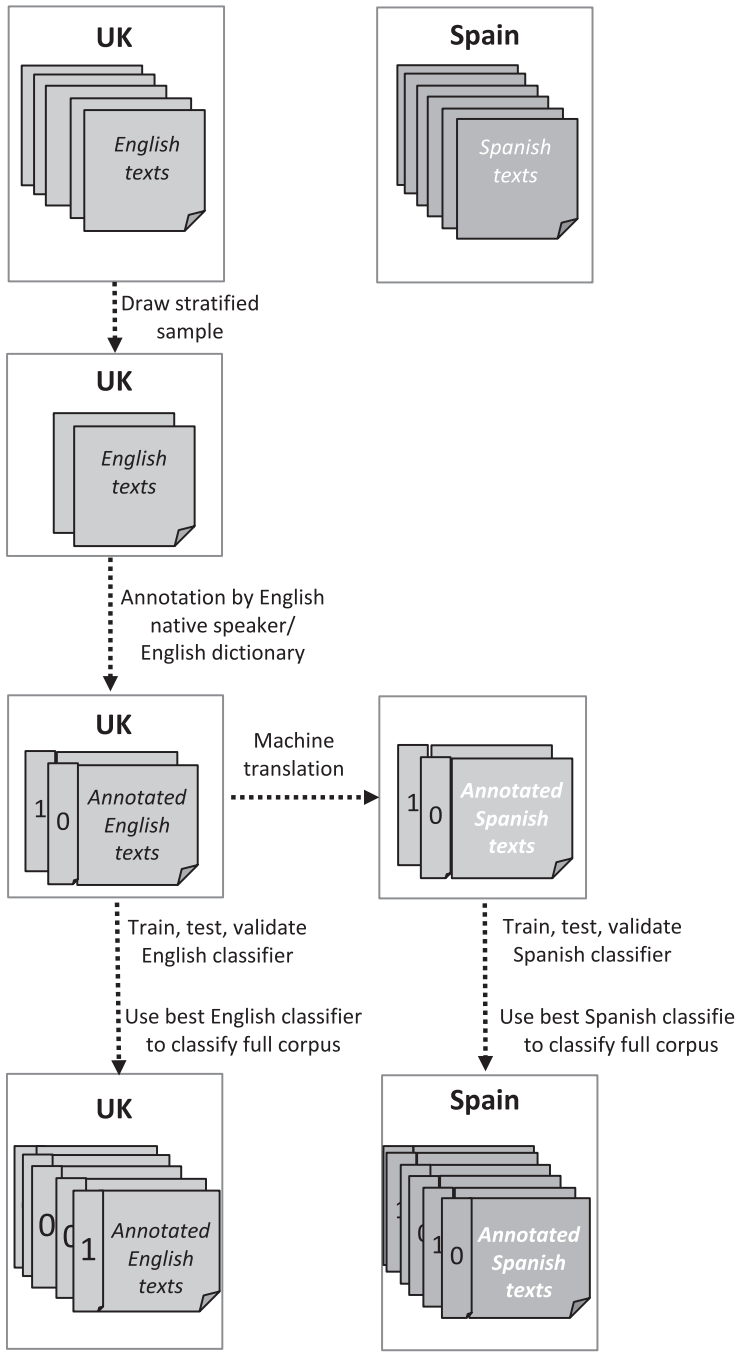
Like strategy B, this strategy builds on the availability or creation of annotated documents in one language. Other than strategy B, where the full corpus is translated, "only" the annotated documents are translated into other languages in strategy C. The annotations are taken along. These newly created data sets are used to train one classifier per language. Examples of this strategy can be found in the computational linguistics literature. Balahur and Turchi (2014) drew upon an English language corpus, manually annotated for sentiment, and used machine translation to create an annotated Spanish, French, and German corpus (see also Banea et al. (2008) for another example.


UK
English texts
Spain
Spanish texts
Draw stratified sample
Draw stratified sample
UK
English texts
Spain
Spanish texts
Annotation by English native speaker/ English dictionary
Annotation by Spanish native speaker/ Spanish dictionary
UK
1
0
Annotated English texts
Spain
1
1
Annotated Spanish texts
Train, test, validate English classifier
Train, test, validate Spanish classifier
Use best English classifier to classify full UK corpus
Use best Spanish classifier to classify full Spanish corpus
UK
0
1
Annotated English texts
Spain
1
0
Annotated Spanish texts

Strategy A


UK
English texts
Spain
Spanish texts
Draw stratified sample
UK
English texts
Machine translation
Annotation by English native speaker/ English dictionary
Spain
English texts
UK
1 0
Annotated English texts
Train, test, validate English classifier
Use best English classifier to classify full translated corpus
Use best English classifier to classify full corpus
UK
0 0 1
Annotated English texts
Spain
1 0
Annotated English texts

Strategy B



Strategy C

Figure 1 Three Strategies for Multilingual Text Classification with SML

Note. For clarity, we show only two countries/languages and one category annotated with either 0 or 1.

Before we get to a step-by-step presentation of our proposed strategy, we briefly evaluate strategies A, B, and C by reviewing the implications of the methodological decisions that concern the selection of documents for annotation, the use of machine translation, and the resources related to annotation and translation.

Selection of Documents for Annotation

If used for comparative research, the key problem of strategies B and C is that classifiers' learning experience is limited to one subset of the originally multilingual corpus, namely the documents of one language. Training an algorithm with data from one national context and subsequently using it to annotate data from another national context may lead to low performance (see the study by Loftis & Mortensen, 2020). As Shi et al. (2010) explain, "similar to domain adaptation in statistical machine learning, due to the discrepancy of data distribution between the training domain and test domain, data distribution across languages may vary because of the difference of culture, people's interests, linguistic expression in different language regions" (p. 1057). This means, in consequence, that the feature sets encountered with the new documents are likely to show differences with the learning feature set and thus contain features where the classifier will have blind spots. If given the choice, it is therefore preferable to prepare training material with document samples for each language and case (i.e., country) of the multilingual corpus.

Use of Machine Translation

Machine translation was shown to be a useful strategy for dealing with scarce resources for multilingual content analyses (De Vries et al., 2018; Lucas et al., 2015), which gives us reason to assume that combining SML with machine translation, as strategies B and C do, is, in principle, a reasonable approach. Still, a loss of quality, due to translation errors, albeit moderate, must be expected, as Balahur and Turchi (2014) show. Since most classifiers in strategy C are trained with machine-translated data, possible translation errors may impact the learning experience of the classifiers. In strategy B, the learning experience is undisturbed from translation errors, since the only classifier learns the relationship between features and classes based on untranslated features. Translation errors become, however, a potential problem when the classifier is used to predict classes for the translated parts

of the full data set. Even if machine translation has matured, if given the option, it appears preferable to train the algorithms with original-language documents and annotate untranslated documents to simply avoid potential translation errors from the onset.

Resources: Annotation and Translation Costs

Considering the two previous sections, some advantages of strategy A are evident: Since the annotation is done for a sample of untranslated documents per language, both language and case context can be taken into account in the annotation process, translation errors are avoided, and algorithms can work with untranslated feature sets. However, strategy A is resource-intensive when it comes to annotation. In this strategy, the languages are processed through different pipelines. In SML, this means that as many different classifiers are trained as there are languages. It is thus necessary to create training data for each language. To establish (valid) comparability of the results across languages and across countries, when designing their studies researchers must pay attention to sample, construct, measurement, and procedural equivalence (Esser & Vliegenthart, 2017; Rössler, 2012).

In respect to the annotation process, sample equivalence means that the documents sampled per case (and here per language) are ideally equivalent across cases (e.g., by selecting articles from the most widely distributed media outlets per country; Rössler, 2012). Construct equivalence relates to a shared understanding or interpretability of the construct to be studied across cases (Boomgaarden & Song, 2019); the construct is described in a codebook used for annotation. Among the different approaches to measurement equivalence (see, e.g., Livingstone, 2003) is one that builds on joint coding training in one project language for all native speakers but involves annotation in the respective native languages (Esser & Vliegenthart, 2017). For automated annotation, such an understanding of measurement equivalence refers to the selection of functionally equivalent keywords (they can be case- and language-sensitive but seek to have the same meaning across cases in respect to the more abstract concept) for the multilingual dictionary. Procedural equivalence means that the annotation procedure must be orchestrated or synchronized as parallel as possible (Rössler, 2012); the reliability of manual annotation is, for example, evaluated across languages, not just for one language (Peter & Lauf, 2002). All in all, the annotation costs involved may explain why hardly any project built annotated document sets for several languages in parallel.

Strategies B and C have one common answer to this problem. They need annotated documents only in one language, which reduces annotation

costs tremendously. Only human coders or a dictionary for one language are required. In the case of strategy B, it is also true that all documents, since they have been completely translated into one language, can be pre-processed with the same methods, which is advantageous for the demanded procedural equivalence. What both have in common, however, and strategy B even more so, are possible costs for the automated translation. When relying on the state-of-the-art translation services provided by large tech companies like Google, Microsoft, or Amazon, the costs may quickly exceed the budget for the large volumes of documents required for SML.[6] We conclude that the most resource-saving scenario seeks to perform coding in one language and reduces machine translation as far as possible.

Introducing Our Methodological Approach

With our methodological approach, we design a strategy, henceforth called strategy D, that allows SML classification for multilingual corpora in a form that includes the advantageous but excludes the less beneficial aspects of strategies A–C as far as possible. It provides an answer to the following question: How can tools like automated annotation and machine translation be most effectively (i.e., high-quality) *and* most efficiently (i.e., resource-saving) employed for the classification of a multilingual corpus with SML?

In strategy D, annotation is conducted only in one language. Unlike strategy A, the documents are machine-translated into this language in advance because this reduces the resources for annotation enormously. Unlike strategies B and C, the annotated documents are sampled from all languages (and cases), not only from one language, to avoid bias issues. The annotations are then projected to the original-language documents, which can subsequently be used to train one classifier per language. Courtney et al. (2020) followed this strategy partly in their study. In contrast, the authors trained and tested the classifier with translated data, and the annotations obtained for the translated documents were not projected to the untranslated equivalent documents. In turn, the advantage of strategy D is that the documents and feature sets used ultimately by the classifier(s) to learn remain untranslated. The classifiers' training is thus done again with language-specific feature sets, which avoids (expensive) full-corpus translation and potential performance loss due to translation errors. Please see Figure 2 for a visualization of strategy D.

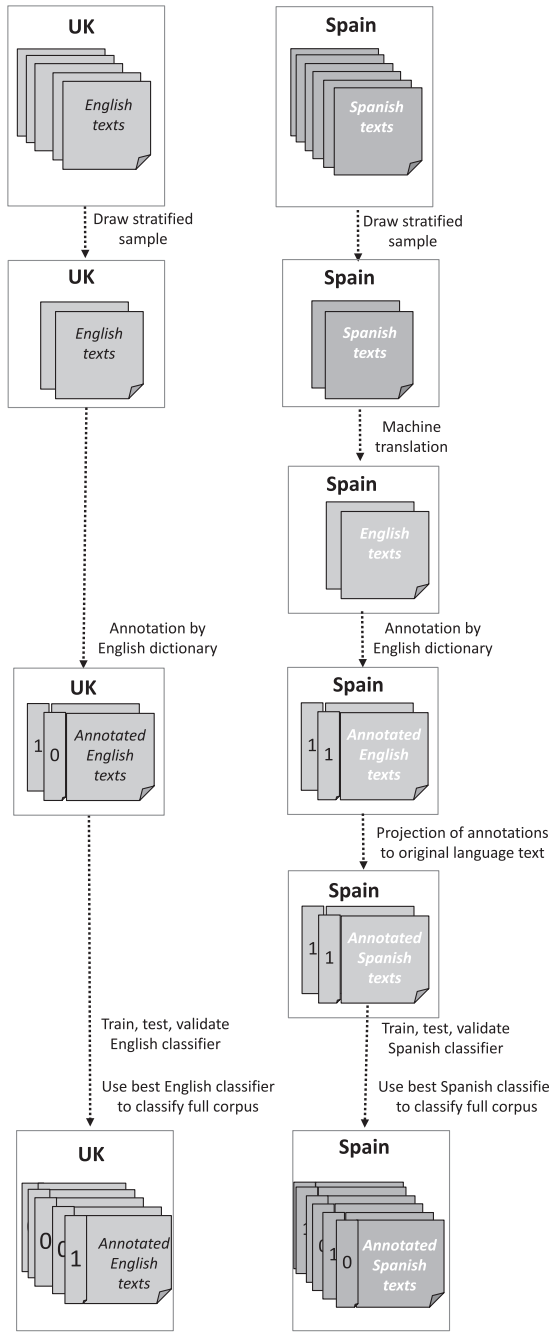


Figure 2 Methodological Setup Strategy D

Strategy D: A Walkthrough

To introduce strategy D, we use the case of European media discourses about migration from seven European countries published in the respective official languages of the countries. Consisting of a total of *N* = 138,388 print and online news articles, this corpus was put together by the authors gathering articles from archives and online news sites. It includes articles from Germany (DE; 46,360), Hungary (HU; 19,704), Poland (PL; 9,601), Romania (RO; 5,571), Spain (ES; 14,646), Sweden (SV; 14,595), and the UK (EN; 27,911) published between January 2017 and November 2018. For more information on included media sources as well as the article selection process and its validation, see Online Appendix Tables A1–A3[8] and in Heidenreich et al. (2020). We investigated four frames frequently recurring in recent literature (e.g., Eberl et al., 2018), namely the economy & budget frame, the labor market frame, the security frame, and the welfare frame. Aiming at concept equivalence, we defined the concepts of the migration frames jointly, based on migration literature and including feedback rounds from experts with expertise for the countries and languages. This way, we tried to establish concept definitions which are located on a level that seek to include all cases and may take into account both transnational discourses as well as country-specific sub discourses. In order to introduce strategy D we decided to present a rather simple classification task, a binary classification task which aims at the differentiation between the presence and absence of a frame. We included four frames instead of one frame only to evaluate whether strategy D is limited to a specific category or also useful for other categories. For detailed definitions of the frames, see Online Appendix Table A4.

Step 1: Selection of corpus samples and their machine translation into one language

We first drew a stratified sample (i.e., every second article by outlet per country) of the full multilingual corpus, obtaining a total of 68,017 articles (DE: 23,101; HU: 9,762; PL: 4,782; RO: 2,690; ES: 7,295; SV: 6,536; EN: 13,851). All non-English articles of this sample were machine-translated into English using the machine translation API from Google (https://cloud.google.com/translate/docs/). This selection enabled us to ensure we end up with a reasonable share of investigated frames compared to the whole corpus and provided the foundation for the further steps of our approach.

Step 2: Annotation of the monolingual document samples with a monolingual dictionary

In a second step, four English dictionaries were applied to the translated, now English version of the articles to unveil the occurrence of the four frames. The dictionaries were carefully designed and validated specifically for the case examined here. Striving for measurement equivalence, we sought to identify dictionary keywords that point to the designed frame concepts and that are functionally-equivalent keywords across cases. We thus selected keywords from available dictionaries for similar concepts and from annotated case-specific party manifestos to include case-specific vocabulary. Case experts reviewed the keyword lists per frame to evaluate their usefulness for the measurement per case. The dictionaries were then refined via a comparison with manually annotated news documents and – once a certain recall and precision was reached – finally validated with another set of manually annotated documents. Human coders annotated the documents based on the frame concept definitions mentioned above. More information about the dictionary creation process, its refinement, and validation can be found in the Online Appendix Tables A5–A7 and in Heidenreich et al. (2020). To allow for matches between the lemmatized keywords of the dictionaries, we lemmatized the English documents using the R Package *UDPipe* (Wijffels, 2019). After the dictionary classification, we knew for each article whether a frame was present (= 1) or not (= 0). A frame was present if at least one frame-specific keyword that referred to the topics of economy, labor, welfare, or security, appeared together with a migration-specific keyword in the same sentence (See Online Appendix pp. 10-11 for more details).

Table 1 shows the dictionary annotation frequencies for each category and language.

Table 1 Dictionary Annotations per Frame and Language (Country)

	DE (Germany)	HU (Hungary)	PL (Poland)	RO (Romania)	ES (Spain)	SV (Sweden)	EN (UK)
Frames	% (document counts)						
Economy & budget	12 (2,668)	18 (1,745)	16 (782)	16 (419)	17 (1,274)	15 (976)	14 (1,996)
Labor market	16 (3,611)	13 (1,239)	17 (815)	14 (390)	15 (1,107)	15 (966)	18 (2,553)
Security	27 (6,134)	38 (3,746)	21 (995)	25 (676)	32 (2,303)	28 (1,821)	30 (4,109)
Welfare	17 (3,904)	10 (989)	11 (511)	10 (269)	14 (1,036)	19 (1,287)	15 (2,110)
Articles (n)	23,101	9,762	4,782	2,690	7,295	6,536	13,851

Step 3: Annotation projection to the original-language version
For each article, we then transferred the dictionary annotations from the translated English version of an article to the original-language version of that same article. For example, a Polish article was classified as "economy" related if the English translation of this article was classified as "economy" related by the English-language dictionary. This projection was central because it allowed us to use the annotated original-language articles as input for the SML, as we want the algorithm to learn from patterns in the original-language documents. The dictionary annotations are therefore considered the "ground-truth" variables (Pilny et al., 2019, p. 4).

Step 4: Classifier selection
The next step included the training and testing of text-classification algorithms with the original-language articles and the projected annotations from step 3. As general pre-processing steps, all original-language articles were lemmatized with the R Package *UDPipe* (Wijffels, 2019), changed to lowercase, and transformed into the tf-idf format. To select the most useful settings for text classification, we compared the performance of different classification approaches while varying common setups per language and frame. Our objectives were mainly to determine what type of algorithm can achieve high performance with the dictionary annotated test data and to assess how performance varies by different sizes of the training and test sets. "Performance," here, relates to F1 scores. F1 is defined as the harmonic mean of recall and precision.[7] All approaches were implemented using the Python packages *Scikit-Learn* (Pedregosa et al., 2011), *Tensorflow* (Abadi et al., 2016), and *Keras* (Chollet, 2018).

To find the best classification approach per frame and language, we used a grid search as (hyper)parameter optimization method. We calculated model performance for three different classifier algorithms, using various combinations of (hyper) parameter values for each of them. The three selected three algorithms are popular in communication research and deemed useful for text-classification tasks (e.g., Burscher et al., 2014; van Atteveldt et al., 2021). First, we implemented *Random Forest* (RF) as a decision tree-based classifier (Breiman, 2001). For the second algorithm, we decided in favor of a *Support Vector Machine* (SVM), a commonly used machine learning approach for two-group classification problems (Cortes & Vapnik, 1995). Finally, we applied a *Multi-Layer Perceptron* (MLP), a neural network architecture (Goldberg, 2017), as a third approach. The selection of MLP and its implementation is guided by a Google text-classification

guide (https://developers.google.com/machine-learning/guides/text-classification/step-4). When comparing different hyperparameters per algorithm we also varied the number of selected features (Kao & Poteet, 2007, p. 175). We used a 3-fold cross-validation approach (Japkowicz & Shah, 2011, p. 163) and random undersampling (Galar et al., 2011). Random undersampling meant that we used the maximum number of positive instances per class (see document counts in Table 1 per frame and language) and randomly sampled the same number of instances for the negative class. The total number of calculated models in this initial selection step was N = 11,088. Please see Online Appendix pp. 11-13 for details on the hyperparameter selection step, 3-fold cross-validation, and random undersampling.

Following this procedure, we selected the optimal number of selected features and the best performing (hyper)parameter configuration per algorithm for each frame and language. Working with this set-up, we again calculated multiple models, this time varying the algorithm and the numbers of training documents for the positive class per classifier per frame and language. To compare the performances of different quantities of data, we applied random undersampling up to 14 times per language and class and created data sets with {100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1,000} instances of the positive class and the respective same number of instances of the negative class.[9] We used a 3-fold cross-validation approach and 3 times re-sampling (Japkowicz & Shah, 2011, p. 163) to ensure robust results. Ultimately, the total number of models considered was N = 9,423 (up to 14 different numbers of documents for the positive class, depending on frame and seven languages, three algorithms, and 3 times re-sampling with 3-fold cross-validation).

Contrasting the performance of the calculated models, we find that while the curve representing the achieved F1 scores steeply increases from 100 to 200 documents for the positive class observations, it flattens shortly after and even slightly decreases following a threshold of 250 to 350 documents, depending on the classifier. This pattern emerges across all algorithms, with varying intensity (see Figure 3). Contrasting the different algorithms, on average, we find that RF (F1: M = .74, SD = .05) and SVM (F1: M = .76, SD = .05) performed similarly well. In turn, results of the performance evaluation with the test sets show that MLP clearly outperforms the other two (F1: M = .93, SD = .09). This is consistent across the four frames as well as the seven languages under investigation.

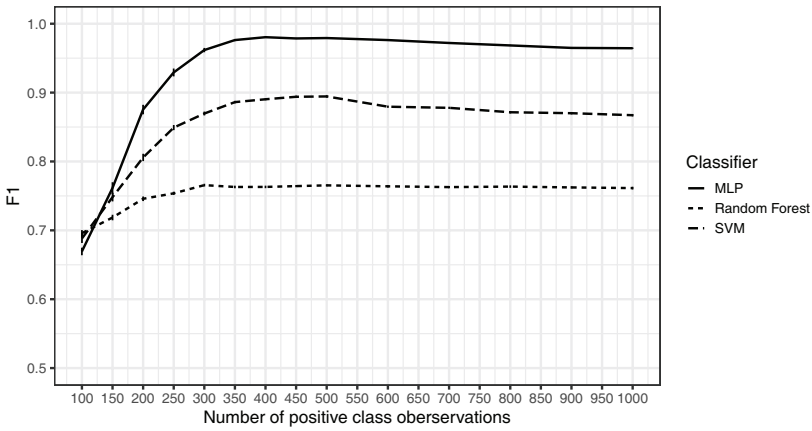


Figure 3 F1 Scores for Different Numbers of Positive Class Observations by Classifier Type

Considering the four frames with respect to average F1 scores, we furthermore see that none of the categories measured performed tremendously better or worse than any other (see Table 2). Lastly, a similar pattern can be described concerning the seven languages examined; algorithms performed equally well across all languages (see Table 2).

Table 2 F1 Scores by Frames and by Languages (Countries)

	DE (Germany)	HU (Hungary)	PL (Poland)	RO (Romania)	ES (Spain)	SV (Sweden)	EN (UK)
Frames	Average F1 Scores (SD)						
Economy & budget	.85 (.11)	.82 (.11)	.84 (.10)	.81 (.09)	.83 (.10)	.83 (.09)	.84 (.11)
Labor market	.83 (.12)	.83 (.13)	.87 (.09)	.84 (.10)	.84 (.11)	.87 (.08)	.86 (.11)
Security	.84 (.11)	.84 (.09)	.85 (.10)	.81 (.10)	.84 (.10)	.82 (.12)	.87 (.08)
Welfare	.83 (.13)	.85 (.12)	.86 (.12)	.80 (.12)	.83 (.11)	.83 (.12)	.86 (.10)

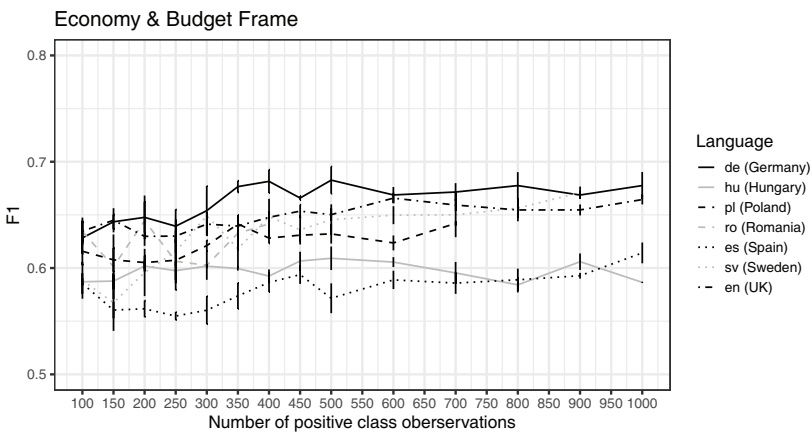
Note. Number of calculated F1 scores = 9,423.

Step 5: Additional evaluation of classifiers with separate manually annotated test data

We incorporated an additional step to assess the strategy to train with dictionary annotated data. To do so, we evaluated the performance of the best performing algorithm (MLP) against the classification decisions of

human coders (Pilny et al., 2019). Thus, not a subset of the dictionary annotated data but separate manually annotated data served as test data. The human-annotated benchmark is a randomly selected sample of migration-related news articles[10] annotated by seven native speakers, who coded the original-language version of the articles. The coders participated in joint coder training to establish a common understanding of the four frame concepts. Intercoder reliability was assessed.[11] The manually annotated test data set included n = 925 migration-related articles per language (and country) and human annotations for the four frames (see Table A7 for annotation frequencies per category and language). We calculated 1,047 models in total (up to 14 different numbers of documents for the positive class, depending on frame and seven languages, 3 times re-sampling). The pre-processing for the manually annotated articles was identical (lemmatization, lowercase, tf-idf transformation) to the pre-processing of the dictionary annotated documents.

We find that MLP achieves satisfactory F1 levels (see Figure 4) throughout most frames and languages also in the manually annotated test set, overall (F1: M = .64, SD = .04). In contrast to the performance of the classifiers on the dictionary annotated test data (see Figure 3), we see that performances in the manually annotated test data steadily improve with increasing size of dictionary annotated training data. While this emerges as an overall pattern, exceptions can be found, such as with the Swedish welfare frame annotations, where the F1 scores remain around the same level (see the graph in the bottom right of Figure 4).



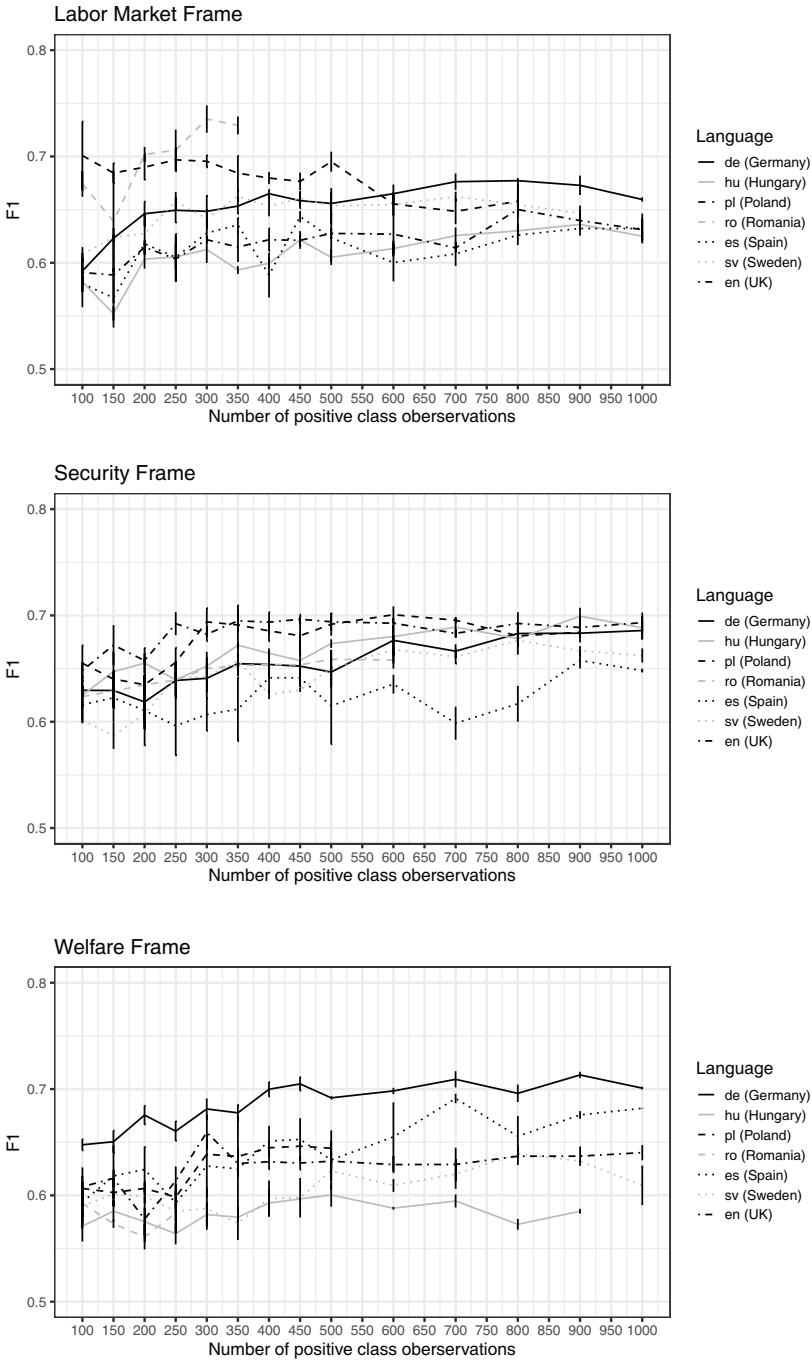


Figure 4 Evaluation of MLP Classifiers Trained with Dictionary Annotated Data Against Manually Annotated Benchmark

Note. The F1 scores are depicted for different numbers of positive-class observations per frame and language (country). Number of calculated F1 scores: $N_{\text{EconomyBudget}} = 261$, $N_{\text{LaborMarket}} = 261$, $N_{\text{Security}} = 279$, $N_{\text{Welfare}} = 446$.

To achieve the best possible measurement for each case when annotating the full corpus, we would use the maximum possible number of positive-class observations per language and category.

Discussion

Multilingual text classification was named one of the "current hottest topics" (Mirończuk & Protasiewicz, 2018, p. 50) in automated text classification. Making a methodological contribution to this field from the perspective of communication research, we introduced an approach that employs SML for the classification of a multilingual news article corpus for comparative communication research. Our contribution here relates primarily to the demonstration of a strategy to obtain annotated documents in sufficient quality and quantity, which is an even greater challenge in the multilingual case than it would be in the monolingual case already.

Our strategy has a number of clear advantages: First, sampling comparable documents from each language for the annotation ensures that classifiers are trained based on material that is representative for the respective case context. Second, the fact that the annotations are projected back to the original-language documents allows to train classifiers with untranslated data and thus undisturbed by potential translation errors. Third, the strategy requires only the means to perform annotation in one language, and the translated subsamples used for annotation can be relatively small (as shown in step 4 and for the specific classification task at hand about 250 to 350 documents per class per language), which is related to the use of comparably few resources required for coding and machine translation (12–16 euros per language when using the Google API standard rates and with an average document length of 11,755 characters).

All in all, with strategy D, we proposed a method to classify a multilingual corpus in a way that is as resource efficient as possible without compromising too much on quality. We showed that the annotation of translated samples with monolingual dictionaries enables the development of good text classifiers for the classification of a large-scale multilingual corpus. An additional test, where the classifiers' predictions are contrasted with manual codings, demonstrates their decent performances across languages and categories.

The application of strategy D for comparative communication research warrants a critical reflection on *construct and measurement equivalence* (Esser & Vliegenthart, 2017; Rössler, 2012). Shifting the annotation process to a monolingual setting, as we and others (e.g., Banea et al., 2008; Courtney et al., 2020) did, is comfortable from a resource perspective. Still, to be useful in a comparative setting, similar to a monolingual annotation setting, procedures have to be in place to ensure that the construct and its measurement are not oriented too much to the single language and case. One approach is to describe the construct on an abstract level that tries to include all possible cases. If human coders "measure", they are ideally trained jointly and have domain knowledge of all cases. The dictionary keywords may be in English (or in another language), but they still reflect ideally all case-specific meanings that are relevant to the more abstract construct. See Online Appendix A4 and pp. 7–11 for our attempt to deal with the challenge of construct and measurement equivalence in a monolingual annotation setting. When it comes to *procedural equivalence* (Rössler, 2012), shifting the annotation process to a monolingual setting appears useful to streamline the coding process. After annotation, with strategy D, the methodological process continues language-specifically. Training several classifiers, one per language, procedural equivalence is at least approached by selecting the same pre-processing steps (e.g., lemmatizing and not stemming, top K feature selection) for all languages. Their execution is then language-specific, in contrast to projects that train a classifier based on translated documents (Courtney et al., 2020) and that can pre-process all documents through the same pipeline. Therefore, if implemented well, strategy D is a good solution from equivalence perspectives.

As evident, strategy D also is not entirely without resource expenditure. We used dictionaries for annotation. The quality of the results thus, of course, strongly depends on the quality of the dictionaries (see Online Appendix A6). We are aware that such dictionaries are not always available in abundance; hence, researchers might often have to weigh the costs to construct such dictionaries. To construct new dictionaries that enable equivalent measurements across cases, close collaboration with language and case experts is strongly recommended. For projects where this does not seem viable, there is the possibility to annotate the translated monolingual documents manually (for a successful test of such a strategy, see Courtney et al., 2020). Besides the costs required for annotation, if annotation is performed (automatically or manually) based on monolingual material, not all documents but at least parts have to be translated, which is likely to involve some costs (for ideas to save on translation costs, see Footnote 5).

In conclusion, while we consider the classification performance all in all sufficient, we do not hide the fact that we observe a drop in performance (variance error) when we apply the classifiers to a manually annotated new test data set. Among the possible reasons for this loss in performance is a vocabulary deviation between the annotated material used for training and the documents of the manually annotated test set. The dictionary annotated data may miss features that occur in the manually annotated test set due to the fact that the publication period of the data sets differs in parts (see Footnote 9). The most obvious solution is to collect more annotated documents of high(er) quality. Before further dictionary improvements or manual annotation are considered, one (comparably simpler) alternative may be worth testing. The minority class in unbalanced annotated data sets may be augmented via forth-and-back translation with machine translation software (Kothiya, 2019). A not yet perfect machine translation service or one that offers multiple translations for certain words may help to produce synonyms for some features and effectively increase the number of annotated documents. Other common ways to improve the learning experience of the classifiers include alternative feature presentation (word embeddings instead of the bag-of-word approach; Chan et al., 2020) or re-sampling strategies (Japkowicz & Shah, 2011).

Finally, we emphasize that although strategy D's performance was somewhat stable across the selected categories and languages, this is, of course, not a free pass for its application to any languages or concepts. The machine translation quality of other language pairs may vary and should be checked, and as expressed elsewhere (e.g., Song et al., 2020; van Atteveldt et al., 2021), the strategy should be carefully validated before being used in a new context and for other languages and concepts.

Supplemental Materials

The Supplemental Materials could be found at:
https://dx.doi.org/10.17605/OSF.IO/VUMH5
https://github.com/Christoph/MultilingualTextAnalysis

Funding Note

This work was supported by Horizon 2020 Framework Programme [grant number 727072].

Noten

1. This work was supported by Horizon 2020 Framework Programme [grant number 727072]. We have no conflict of interest to disclose. Correspondence concerning this article should be addressed to Fabienne Lind, Kolingasse 14-16, 1090 Vienna, Austria, Email: fabienne.lind@univie.ac.at
2. https://github.com/Christoph/MultilingualTextAnalysis
3. Different languages frequently relate to the different cases, which are often (but not always) countries (e.g., comparison of language regions in Switzerland or Belgium).
4. The approach to transfer a multilingual data set into a monolingual version prior to analysis to then work with monolingual instruments and methods was outlined in the seminal paper by Lucas et al. (2015).
5. To improve the error estimation procedure of this test, the annotated data are usually not simply split once into training and test data (holdout method), but "re-used" in the form of re-sampling techniques, such as repeated k-fold cross-validation (Japkowicz & Shah, 2011, p. 163).
6. For example, the translation of 1 million characters costs $20 with the Google API. To put this number in perspective, in our sample, the average article length is 11,755 characters, so 1 million characters corresponds to the translation of 85 articles. We like to point to the cost-saving options like free start contingencies and research grants offered by the mentioned tech companies.
7. https://dx.doi.org/10.17605/OSF.IO/VUMH5
8. F1= (2 * recall * precision) / recall + precision
9. Due to data availability (see Table 1), the maximum number of positive-class documents for Polish documents was 700 for the economy & budget frame, 800 for the labor market frame, and 900 for the security topic. For Romanian, the respective maximum numbers were 400 documents (for economy & budget), 350 (for labor market), and 600 (for security). For Sweden, the limit for the labor and the security categories was 900.
10. The sample was drawn from a corpus of migration-related articles selected based on the same search strings as the corpus introduced on page 17 but includes news articles published earlier, namely between January 2000 and December 2017.
11. All seven coders classified 70 English (untranslated) articles (Krippendorff's alphas: .71–.79). In addition, every native speaker manually coded 50 untranslated articles (in the respective native language). These annotations were then compared with the annotations of a native English speaker, who annotated the same articles but in their translated version (Krippendorff's alphas: .64–.92).

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … & Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16) (pp. 265-283).

Alkhair, M., Meftouh, K., Smaïli, K., & Othman, N. (2019, October). An arabic corpus of fake news: Collection, analysis and classification. In K. Smaïli (Ed.), *Proceeding of the 7th Conference on Arabic Language Processing* (pp. 292-302).

Baden, C., Schonvelde, M., Pipal, V., & an der Velden, M., (2020). Three gaps in computational methods for social sciences: A research agenda. (Working paper).

Baden, C., & Stalpouskaya, K. (2015). *Common methodological framework: Content analysis. A mixed-methods strategy for comparatively, diachronically analyzing conflict discourse*. INFOCORE Working Paper 2015/10. www.infocore.eu/results/working-papers/

Balahur, A., & Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1), 56–75. https://doi.org/10.1016/j.csl.2013.03.004

Banducci, S., de Vreese, C. H., Semetko, H. A., Boomgarden, H. G., Luhiste, M., Peter, J., … & Xezonakis, G. (2014). European Parliament election study, longitudinal media study 1999, 2004, 2009. GESIS Data Archive, Cologne. ZA5178 Data file Version 1.0.0. doi:10.4232/1.5178

Banea, C., Mihalcea, R., Wiebe, J., & Hassan, S. (2008). Multilingual subjectivity analysis using machine translation. In M. Lapata & H. Tou Ng (Eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 127–135). Association for Computational Linguistics. https://www.aclweb.org/anthology/D08-1

Baumgartner, F. R., Breunig, C. & Grossman E. (2019). *Comparative policy agendas: theory, tools, data*. Oxford University Press. 2441/2pt196maou9nurq8q8eudu9p2i

Boomgaarden, H. G., & Song, H. (2019). Media use and its effects in a cross-national perspective. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 71(1), 545–571. https://doi.org/10.1007/s11577-019-00596-9

Boot, P. (2021, January 12). Machine-translated texts as an alternative to translated dictionaries for LIWC. (Working Paper). https://doi.org/10.31219/osf.io/tsc36

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

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

Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., & De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised

machine learning approaches to frame analysis. *Communication Methods and Measures*, 8(3), 190–206. https://doi.org/10.1080/19312458.2014.937527

Caviedes, A. (2015). An emerging 'European'news portrayal of immigration? *Journal of Ethnic and Migration Studies*, 41(6), 897–917. https://doi.org/10.1080/1369183X.2014.1002199

Chan, C. H., Zeng, J., Wessler, H., Jungblut, M., Welbers, K., Bajjalieh, J. W., ... & Althaus, S. L. (2020). Reproducible extraction of cross-lingual topics (rectr). *Communication Methods and Measures*, 14(4), 285–305. https://doi.org/10.1080/19312458.2020.1812555

Chang, C., & Masterson, M. (2020). Using word order in political text classification with long short-term memory models. *Political Analysis*, 28(3), 395-411.

Chollet, F. (2018). Keras: The python deep learning library. *Astrophysics Source Code Library*, ascl-1806.

Cohen, A. A. (2012). Benefits and pitfalls of comparative research on news: Production, content, and audiences. In I. Volkmer (Ed.), *The handbook of global media research* (pp.533–27). Blackwell Publishing Ltd. doi:10.1002/9781118255278

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Courtney, M., Breen, M., McMenamin, I., & McNulty, G. (2020). Automatic translation, context, and supervised learning in comparative politics. *Journal of Information Technology & Politics*, 17(3), 208-217. https://doi.org/10.1080/19331681.2020.1731245

De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis*, 26(4), 417–430. doi:10.1017/pan.2018.26

Deng, X., Li, Y., Weng, J., & Zhang, J. (2019). Feature selection for text classification: A review. *Multimedia Tools and Applications*, 78(53), 3797–3816. https://doi.org/10.1007/s11042-018-6083-5

Eberl, J. M., Meltzer, C. E., Heidenreich, T., Herrero, B., Theorin, N., Lind, F., ... & Strömbäck, J. (2018). The European media discourse on immigration and its effects: A literature review. *Annals of the International Communication Association*, 42(3), 207–223. https://doi.org/10.1080/23808985.2018.1497452

Esser, F., & Hanitzsch, T. (2012). On the why and how of comparative inquiry in communication studies. In F. Esser & T. Hanitzsch (Eds.), *Handbook of comparative communication research* (pp. 3–22). Routledge. https://doi.org/10.4324/9780203149102

Esser, F., & Vliegenthart, R. (2017). Comparative research methods. The international encyclopedia of communication research methods, 1-22. https://doi.org/10.1002/9781118901731.iecrm0035

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1), 1-309. https://doi.org/10.2200/S00762ED1V01Y201703HLT037

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

Gründl, J. (2020). Populist ideas on social media: A dictionary-based measurement of populist communication. *New Media & Society*. Advance online publication. https://doi.org/10.1177/1461444820976970

Heidenreich, T., Lind, F., Eberl, J.-M., Galyga, S., Edie, R., Herrero-Jiménez, B., Gómez Montero, E.L., Berganza, R., & Boomgaarden, H.G. (2020). *REMINDER: Short Term Media Analysis on Migration 2017-2018* (*OA edition*), [Dataset and documentation]. AUSSDA Dataverse. doi: 10.11587/LBSMPQ

Hopmann, D. N., Esser, F., de Vreese, C. H., Aalberg, T., van Aelst, P., Berganza, R., … & Strömbäck, J. (2016). How we did it: Approach and methods. In C. De Vreese, F. Esser, & D. N. Hopmann (Eds.), *Comparing political journalism* (pp. 10–21). Routledge. https://doi.org/10.4324/9781315622286

Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: A classification perspective*. Cambridge University Press. https://doi.org/10.1017/CBO9780511921803

Kananovich, V. (2018). Framing the Taxation-Democratization link: An automated content analysis of cross-national newspaper data. *The International Journal of Press/Politics*, 23(2), 247–267. https://doi.org/10.1177/1940161218771893

Kao, A., & Poteet, S. R. (Eds.). (2007). *Natural language processing and text mining*. Springer Science & Business Media.

Karan, M., Šnajder, J., Širinić, D., & Glavaš, G. (2016, August). Analysis of policy agendas: Lessons learned from automatic topic classification of croatian political texts. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 12–21). https://www.aclweb.org/anthology/W16-2100

Kothiya, Y. (2019). How I handled imbalanced text data. Blueprint to tackle one of the most common problems in AI. Blog post. https://towardsdatascience.com/how-i-handled-imbalanced-text-data-ba9b757ab1d8

Lind, F., Eberl, J. M., Eisele, O., Heidenreich, T., Galyga, S., & Boomgaarden, H. G. (2021). Building the Bridge: Topic Modeling for Comparative Research. *Communication Methods and Measures*. https://doi.org/10.1080/19312458.2021.1965973

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.

Livingstone, S. (2003). On the challenges of cross-national comparative media research. *European Journal of Communication*, 18(4), 477–500. https://doi.org/10.1177/0267323103184003

Loftis, M. W., & Mortensen, P. B. (2020). Collaborating with the machines: A hybrid method for classifying policy documents [Supplemental material]. *Policy Studies Journal*, 48(1), 184–206. https://doi.org/10.1111/psj.12245

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, 23(2), 254–277. http://nrs.harvard.edu/urn-3:HUL.InstRepos:38057808

Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, 106, 36–54. https://doi.org/10.1016/j.eswa.2018.03.058

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Peter, J., & Lauf, E. (2002). Reliability in cross-national content analysis. *Journalism & Mass Communication Quarterly*, 79(4), 815-832. https://doi.org/10.1177/107769900207900404

Pilny, A., McAninch, K., Slone, A., & Moore, K. (2019). Using supervised machine learning in automated content analysis: An example using relational uncertainty. *Communication Methods and Measures*, 13(4), 287–304. https://doi.org/10.1080/19312458.2019.1650166

Proksch, S.-O., Lowe, W., Wäckerle, J., & Soroka, S. (2019). Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, 44(1), 97–131. https://doi.org/10.1111/lsq.12218

Reber, U. (2019). Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication Methods and Measures*, 13(2), 102–125. https://doi.org/10.1080/19312458.2018.1555798

Rössler, P. (2012). Comparative content analysis. In F. Esser & T. Hanitsch (Eds.), *The handbook of comparative communication research* (pp. 481–490). Routledge. https://doi.org/10.4324/9780203149102

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

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

Schuck, A. R., Vliegenthart, R., Boomgaarden, H. G., Elenbaas, M., Azrout, R., van Spanje, J., & De Vreese, C. H. (2013). Explaining campaign news coverage: How medium, time, and context explain variation in the media framing of the 2009 European parliamentary elections. *Journal of Political Marketing*, 12(1), 8–28. https://doi.org/10.1080/15377857.2013.752192

Sebők, M., & Kacsuk, Z. 2020. The multiclass classification of newspaper articles with machine learning: The hybrid binary snowball approach. *Political Analysis*. Advance online publication. doi: 10.1017/pan.2020.27

Shi, L., Mihalcea, R., & Tian, M. (2010, October). Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 1057–1067). https://www.aclweb.org/anthology/D10-1000

Song, H., Tolochko, P., Eberl, J. M., Eisele, O., Greussing, E., Heidenreich, T., … & Boomgaarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*, 37(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752

Strömbäck, J., Andersson, F., & Nedlund, E. (2017). *Invandring i medierna. Hur rapporterade svenska tidningar åren 2010–2015?* [Immigration in the media. How did Swedish newspapers report in the years 2010–2015?] Stockholm: Delmi/Statens Offentliga Utredningar.

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*. Advance online publication. https://doi.org/10.1080/19312458.2020.1869198

Volkens, A., Lehmann, P., Matthieß, T., Merz, N., Regel, S., & Werner, A. (2015). The manifesto data collection: Manifesto project (MRG/CMP/MARPOR, Version 2015a) [Computer software]. Wissenschaftszentrum Berlin für Sozialforschung. https://doi.org/10.25522/manifesto.mpds.2015a

Watanabe, K. (2020). Latent semantic scaling: A semisupervised text analysis technique for new domains and languages. *Communication Methods and Measures*. Advance online publication. https://doi.org/10.1080/19312458.2020.1832976

Wijffels, J., Straka, M., & Strakov, J. (2019). Udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the UDPipe NLP toolkit. (R Package Version 0.6). [Computer software]. Available at https://cran.r-project.org/web/packages/udpipe/index.html

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

About the author

Correspondence address: Fabienne Lind, Kolingasse 14-16, 1090 Vienna, Austria