Dictionary-based and machine learning classification approaches: a comparison for tonality and frame detection on Twitter data

Maud Reveilhac and Davide Morselli

Faculty of Social and Political Sciences, Institute of Social Sciences, Life Course and Social Inequality Research Centre, Lausanne University, Lausanne, Switzerland

ABSTRACT
Automated text analysis methods have made it possible to classify large corpora of text by measures such as frames and tonality, with a growing popularity in social, political and psychological science. These methods often demand a training dataset of sufficient size to generate accurate models that can be applied to unseen texts. In practice, however, there are no clear recommendations about how big the training samples should be. This issue becomes especially acute when dealing with texts skewed toward categories and when researchers cannot afford large samples of annotated texts. Leveraging on the case of support for democracy, we provide a guide to help researchers navigate decisions when producing measures of tonality and frames from a small sample of annotated social media posts. We find that supervised machine learning algorithms outperform dictionaries for tonality classification tasks. However, custom dictionaries are useful complements of these algorithms when identifying latent democracy dimensions in social media messages, especially as the method of elaborating these dictionaries is guided by word embedding techniques and human validation. Therefore, we provide easily implementable recommendations to increase estimation accuracy under non-optimal condition.

ARTICLE HISTORY
Received 5 April 2021
Accepted 11 January 2022

KEYWORDS
Text analysis; dictionary-based approach; machine learning; democracy frames; tonality detection

Textual analysis of social media
Political science and political psychology have looked with increased interest at social media analysis. On these platforms, in real time, people express opinions and evaluations of facts, the content of which can be used to measure peoples’ thoughts and feelings as an alternative to traditional surveys (Schwartz and Ungar 2015). This is typically done relying on text classification methods, which provide a way of estimating a category (or class) to which a given piece of text might belong. Grimmer and Stewart (2013) summarize different content analytical techniques lying on a continuum from dictionary-based, to (semi-)supervised (e.g. machine learning from a labelled training set), and fully data-

CONTACT Maud Reveilhac maud.reveilhac@unil.ch
© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
driven methods (e.g. topic modelling). Because there is no guarantee that unsupervised methods return categories of theoretical interest, we suggest that it is important to develop a methodology that refine text classification combining dictionary and (semi-)supervised approaches.

Though these methods are under constant development, we argue that their application is sometimes limited for political and social scientists because of the narrowness of research questions and the consequent limited amount of data, compared with data science research. However, the baby shouldn’t be thrown out with the bath water. In this article, we provide an example of application of some of these techniques to investigate tonality and frame detection (which are sub-tasks of text classification) in tweets related to democracy. More precisely, we address two main research questions: first, what are the advantages and disadvantages of each approach when relying on a small data set that is imbalanced with respect to the class distribution? And, are partial data-driven techniques for custom dictionary inducement a worthy complement (or alternative) to other off-the-shelf lexical resources and supervised machine learning (ML) techniques in the realm of text classification?

The first research question draws from consideration of the difficulty supervised learning has in classifying the data on small data sets. Small data sets constitute a problem for the performance of machine learning classifiers, especially due to the limited number of examples. This difficulty is even amplified when the distribution of the data is skewed towards a few dominant categories in the data. Although small dataset and skewness of the data are separate issues for supervised learning, their combination can be very demanding for machine learning approaches. Here, dictionary-based approaches are especially beneficial for enhancing external generalizability (Amsler 2020, 135). Indeed, in contrast to supervised models that learn to classify the text into known categories based on a labelled training data set, dictionaries can cover cases that are not represented in the training set. For this reason, several lexicons have been used extensively and on many different domains for classification tasks – such as the General Inquirer from Stone, Dunphy, and Smith (1966), LIWC from Pennebaker et al. (2015) and Lexicoder from Young and Soroka (2012).

The second research question stems from recent advances in the methodology of deriving lexical resources where word embeddings are a component of the dictionary induction and enrichment (see Hamilton et al. 2016; Amsler 2020). This strategy has the advantage of increasing the comparability to other data-driven approaches. We also insist on the fact that an iterative procedure between a data-driven step and a human validation step is paramount for allowing the inclusion of as much relevant information as possible in the dictionary.

Our empirical application is conducted on a corpus of tweets related to democracy. For the framing detection task (see also Amsler, Wüest, and Schneider 2016; Wüest, Amsler, and Schneider 2017), we aim to detect the presence (or absence) of democracy dimensions defined on the basis of previous survey research (Kriesi et al. 2013; Ferrín 2018). Here, we define frames as ‘schemata of interpretation’ (Goffman, 1974, 21) referring to the description (and the interest manifested) of the democratic decision-making process mentioned in tweets (e.g. political responsiveness or accountability). For the tonality detection task, we aim to detect the overarching sentiment conveyed in a given tweet, that is, its positivity or negativity. For both tasks, we compare the classification
accuracies from a custom hand-curated dictionary, off-the-shelf lexicons, and supervised ML models.

Our objective is not to build ‘state-of-the-art’ classifiers with optimal performance, but to understand how validly each method classifies the data under suboptimal conditions, that is utilizing a small and skewed training set. Thus, our goal is not to demonstrate that a classification method will always provide better classification accuracy than another. Instead, we aim to point how to test several methods and to decide which is the best for the given task and data. In this view, we also provide some guidelines to help researchers develop a custom dictionary and pre-process short texts before any tonality and frame detection can be produced.

Established methodologies in text classification

Dictionary-based and supervised learning approaches

When the volume of data is too large to be manually analysed, automated tools are needed to detect sentiment. Neuendorf (2016, 147) has argued that the fully automated approach is not a new procedure since it was introduced more than half a century ago with the General Inquirer (Stone, Dunphy, and Smith 1966), and has been widely used since then.

Dictionaries are the perhaps the most intuitive way of classifying texts according to a priori defined classes. A dictionary uses the rate at which key words appear in a text to classify documents into categories (e.g. tweets, Facebook posts, news articles). It can do it either in a dichotomous manner or by using scores. Either way, the dictionary will correctly identify the categories only if the words contained closely align with how the language is used in the particular context under investigation. Indeed, the application of an ‘off-the-shelf’ dictionary to an area of research outside the substantive domain from which it has been developed can lead to classification errors, especially because similar terms can have different connotations in different contexts (see Loughran and McDonald 2011). The writing conventions of different types of text (e.g. tweets often contain concatenated expressions, such as ‘ClimateChange’) further complicates the generalizability potential of dictionaries. In addition to the potential lack of domain generalizability, another possible pitfall of dictionaries refers to their restrictive domain coverage (or scope). Therefore, although a comprehensive set of keywords mapping unambiguously to a concept can produce highly reliable and efficient results, this method is sometimes criticized for its simplicity, and for its persistent difficulty in achieving completeness. To date, however, there exist methods of mitigating both of these difficulties – namely, the external generalizability and the vocabulary coverage – by making use of word embeddings to expand the dictionaries (e.g. Amsler 2020).

Supervised learning methods provide an alternative for text classification into predetermined categories. They follow the following strategy. First, a sample of the corpus (the training dataset) is coded by humans for tone and frames. Then a classification method (ML algorithm) is selected, and the classifier is trained to predict the manually assigned labels within the training dataset. Here, multiple classification methods are generally applied and tested for minimum levels of accuracy. Finally, the chosen classifier is applied to the entire corpus to predict previously unseen texts (those not labelled by humans).
The supervised methods require it to be demonstrated that the classification from the ML algorithm replicates hand coding. A major advantage of this approach over the dictionary method is that it is necessarily domain specific (Grimmer and Stewart 2013, 275). However, it also has disadvantages, such as the necessity to label a substantial amount of text to train a reliable model (as the algorithm employed only learns from the features present in the training set). It also requires that the training data set not be skewed against a few classes (see Haixiang et al. 2017), otherwise the model may tend to overestimate larger categories at the expenses of smaller and sparser ones. For instance, in the case of tonality detection, most of the words carry no sentiment and the chosen algorithm can learn to wrongly associate neutral words with positive or negative sentiment (e.g. for the sentence ‘e-voting sucks’, the classifier will learn to associate the e-voting with the sentiment negative, and unlearning this will require training set instances with the word e-voting in them that are labelled positive). To optimize supervised models trained on small (or skewed) data sets, one prominent method for text classification is unsupervised pre-training. This approach has been widely adopted with the introduction of pre-trained word embeddings (Mikolov et al. 2013; Joulin et al. 2017), which rely on large data corpora. A recent strong performer in this line of research is BERT (Devlin et al. 2018). BERT models are unsupervised language representation, pre-trained using a plain text corpus. Thus, BERT is useful for generating context-specific embeddings, providing a pre-trained universal model. However, one of its major drawbacks is the computational resources needed to fine-tune and make inferences, especially on imbalanced classes. An additional difficulty with supervised models is that there is no easy way to fix classification errors (e.g. no control over induced learned parameters in the model).

In a nutshell and following Schwartz and Ungar (2015), no method is perfect, and each has advantages and pitfalls. Dictionary-based methods are accessible, theory-driven, abstract, and can be used with small samples, but they may overlook semantic context and misclassify text. Supervised learning can overcome some of these challenges for assigning documents to predetermined categories. Especially, it is necessarily domain specific and provide clear statistics summarizing model performance. However, it need large data sets to ‘learn’ from the data to make accurate predictions.

We argue that the choice of one method over the other needs to be based on their performance, and that is necessarily context specific. As previously noted, some researchers warn against using ‘off-the-shelf’ dictionaries and emphasize the need to adapt them, or even create them from scratch for the task at hand (Grimmer and Stewart 2013, 275). However, the dictionary-based approach enables us to follow a strategy where transparency and sustainability are central pillars. This constitutes a major argument compared to the application of supervised learning approaches, which need a large amount of data to be trained and to balance possible class imbalance characterizing the data. However, the performance dictionary-based methods need to be established on a case-to-case basis.

**Constrained but realistic research scenario: small and skewed manually labelled training set**

A recent study from Barberá et al. (2021) provides guidance for researchers about the steps that need to be taken before any tonality assessment can be produced from newspaper coverage. The authors point out the decisions behind each analytical step, going
from corpus selection to the choice of coding units, the trade-off between the number of annotators and the number of coded documents, as well as the comparison between supervised ML and a dictionary-based approach for tonality detection. They found that ML algorithms outperform dictionaries for tonality detection given a large enough training dataset. Their findings about the better performance of ML methods over dictionaries are congruent with the conclusions from Hartmann et al. (2019) in the marketing domain on social media.

In our study, we follow a similar endeavour in terms of assessing which classification method works best on a training data set. However, our study covers another research scenario that is characterized by the following aspects: we rely on tweets as textual data, we use a small sample of annotated data, our data are skewed toward few classes, we focus on tonality and frame detection, we build a custom dictionary, which means that we compare three classification methods (custom dictionary, off-the-shelf dictionary, and supervised learning models).

The research scenario that we propose is not unlikely. For instance, training data for generating novel prediction can be expensive and might not be possible in many real circumstances as supervised classifiers need to be trained on a quite large hand-labelled dataset to have enough predictive power. Furthermore, though social media are frequently used to discuss politically relevant topics, the popularity of some topics is often quite limited (e.g. fairness of the electoral system), meaning that the amount of data pertaining to one category can be underrepresented compared to other classes.

According to the existing literature aiming for classification accuracy, what is considered a ‘small’ training data set typically contains less than 2000 entries (Riekert, Riekert, and Klein 2021), which represent the usual size in social science. This, however, represents a suboptimal scenario when compared to the context of hundreds of thousands of labelled data that is used in computational linguistics and data science to build the ML models (e.g. Zhang, Zhao, and LeCun 2015; Joulin et al. 2017; Shen et al. 2018). The size of the training set is an important factor in determining which classification approach (dictionary or machine learning) or which classification algorithm is ultimately used. For instance, concerning ML, the models performing well on large training sets sizes tend to neglect the effect of the classes with very few examples on the performance (Cortes et al. 1994). Barberà et al. (2021) further discuss the trade-off between maximizing the number of coded documents by single coders to increase the size of the training set or having fewer annotated documents by multiple coders to increase the coding reliability. Their results showed that the ‘informational gains from increasing the number of documents coded are greater than from increasing the number of codings of a given document’ (30). Different research goals therefore imply different methodological strategies. Thus, researchers need to make the theoretically and practically appropriate choices in terms of the methods applied.

Sentiment analysis and frame detection as classifications tasks

Tonality and frame detection are two important research areas of text classification. They have become increasingly prevalent in research fields where texts are the primary data source. For instance, over 1200 papers applying automated text classification of sentiment have been published to date, spanning organizational science and marketing, psychology,
This line of research has doubled in size in the last three years, showing an expanding importance of these methods in different disciplines.

Tonality detection is often referred as sentiment analysis (or opinion mining) and is conceived as a flexible and powerful tool in some branches of political research, such as political psychology or electoral studies. Traditionally, sentiment analysis was conducted using dictionaries of word polarities (Young and Soroka 2012). Although these approaches are being increasingly replaced by supervised classifications using supervised learning approaches, they continue to be actively developed. The long-standing popularity of sentiment analysis can be explained by the availability of ready-to-use methods and software, such as the Linguistic Inquiry and Word Count (LIWC, Pennebaker et al. 2015). These methods allow researchers either to rely on pre-compiled models (e.g. tonality dictionaries) or to easy tune research-specific dimension to be highlighted in text (Iliev, Dehghani, and Sagi 2015).

In parallel, a growing literature has aimed to develop tools for classifying texts in terms of issue or frame categories. Gilardi and Wüest (2018) summarize contributions relying on text applications for policy analysis into three broad research goals: extraction of specific information (concept identification), theory-driven allocation (classification), and inductive exploration of the underlying dimensionality (discovery). While fully automated (or inductive data-driven) methods of document classification such as topic modelling are a more common approach for frame detection (e.g. Blei, Ng, and Jordan 2003; DiMaggio, Nag, and Blei 2013; Roberts et al. 2014; Gilardi, Shipan, and Wüest 2021), there have also been works using dictionary-based approaches. Wüest, Amsler, and Schneider (2017) detected the presence (or claimed absence) of accountability of new forms of governance. In a similar vein, the Comparative Agendas Project aims to compare policies worldwide, thus investigating the trends in policymaking across time and between countries. In the frame of this project, the Lexicoder Topic Dictionaries were created to capture topics in news content, legislative debates, and policy documents (Albaugh et al. 2013). In addition to dictionary-based approaches, supervised algorithms have also helped researchers to identify dimensions from textual data. For instance, the Media Frames Corpus (Card et al. 2016) proposes a dataset of annotated news articles on 15 general-purpose meta-frames (here called ‘framing dimensions’) and can enable researchers to develop and empirically test models of framing. Other examples include research by García-Marín and Calatrava (2018) to classify frames in the media about the refugee crisis and Gilardi and colleague’s study (2021) about the co-evolution of different agendas (namely, the traditional media, parties’ social media, and politicians’ social media agendas) along specific policy issues.

The present study: tonality and frame detection in tweets related to democracy

The focus of this paper is on the evaluation of the tonality and frames of tweets related to democracy. We test the performance of a custom dictionary, off-the-shelf dictionaries, and unsupervised algorithms to achieve these classification tasks. After annotating tonality and theoretically relevant dimensions of democracy in a sample of randomly selected tweets from our corpus, we show how different dictionary-based and supervised classification approaches cope with the task of automatically detecting content and sentiment about democracy. We discuss the conditions under which the different approaches
provides better classification results. Then, relying on the best classification models, we estimate tonality and democracy dimensions for previously unclassified tweets.

Unlike studies relying on large corpora, we are interested here in providing a procedure to obtain reliable results with small and skewed annotated data sets. We therefore draw from empirical works aiming to compare different research methods for text classification (e.g. Rooduijn and Pauwels 2011; Hartmann et al. 2019; Barberá et al. 2021).

Using democracy as the study focus introduces several interesting aspects to our purposes. Democracy is, without doubt, one of the most complex concepts of contemporary political science. It can be framed in different way, with some frames (e.g. representative-ness, responsiveness) being less spares than others (e.g. sovereignty; Fishman 2016).

Democracy has also been the subject of multiple opinion surveys. To date, social media offer an alternative view on the working of democracy. On social media, discussions about democracy can arise from a variety of stimuli, such as attention to particular events or scandals, and personal motivation to post or not post a tweet. Even if unrepresentative of the wider public (see discussions of social media biases by Japec et al. (2015) and Schober et al. (2016)), social media discussions can thus serve as a complement to more stable surveyed attitudes. Therefore, the last section of our paper is dedicated to the correspondence between attitudes found in surveys and social media prevalence on similar democracy dimensions.

**Methodology**

**Data collection: tweets related to democracy using a list of search-queries**

We built our corpus by retrieving tweets related to democracy based on a list of relevant terms referring to the workings of democracy and extracted from in the main Swiss German³ and French⁴ language newspapers in 2018. Newspapers were retrieved from the Swissdox repository using ‘democracy’, ‘populism’, and ‘Swiss people’ as search entries for article titles. The lemmatized articles’ text was used to extract the top 50 terms associated with the word ‘democracy’. This extracted list was thus used as search keywords in the Twitter API to retrieve tweets. The restriction to three keywords could lead to a specific pre-selection and hence to a bias in the association scores, but the newspaper search only constituted the first step upon which the final list of search-queries is elaborated.

After retrieving the tweets, we extracted the hashtags from the corpus of collected tweets and added the ones related to democracy into the search list. We retrieved tweets with the extended list once again to make sure to include all relevant hashtags. Our final list of search-queries contained 51 French terms, 56 German terms, 50 Italian terms, and 71 hashtags related to general policymaking in Switzerland (see Annex 1: List of search-queries to collect the tweets).

The extracted tweets were also filtered by time and location. With respect to location, Twitter provides two classes of geographical metadata. The tweet location, which is available when users share location at time of tweeting, and the account location, which is based on the ‘home’ location provided by users in their public profile. However, very few users provide these meta-data, which is why we adopted another retrieval strategy. First, we concatenated the search-queries with the word ‘Switzerland’. In this way, we obtain tweets that entail one of our search-queries and mention Switzerland. Second,
we used the possibility of retrieving tweets from users situated in each region by specifying a geographical radius in addition to the search-queries. In this way, we obtained tweets that entail one of our search-queries and that are posted by users mainly in Switzerland or in regions very close to it. In a final step, we kept only the tweets emitted from January to December 2018, and collected the replies to those tweets to increase the size and variability of the corpus. Our final corpus of tweets is composed of 296,375 German, French, and Italian tweets.

Translation and pre-processing steps

To overcome the multilingual diversity in Switzerland, we decided to translate every tweet into English using Google Translate (see pre-processing section below). There are additional reasons for translating the tweets into English. Firstly, we should generate a sufficient amount of data to train ML models (see section 3.6). Secondly, we should have a sufficient amount of data to amplify our custom dictionary (see section 3.4). Thirdly, and more pragmatically, most of the off-the-shelf dictionaries are exclusively in English. Finally, Google Translate works best when translating into English, thus homogenizing the translation biases.

The translation process nonetheless entails two essential limitations. The first is linked to the interpretive sophistication. Indeed, words that pertain to complex socio-political phenomena can evoke different meanings regardless of the translation, a particularity with which the translator might not always cope. For instance, the word ‘nation’ can be translated to ‘state’ or ‘republic’ depending on the original language and on the context known by the translator. The second limitation is linked to the vocabulary coverage, which might be oriented toward most commonly used words instead of sophisticated expressions.5

Before translation is applied, we conducted several pre-processing steps, namely links removal and split of concatenated words (e.g. ‘#ClimateChange’ becomes ‘Climate Change’). These pre-processing steps were important for maximizing the correctness of the translation. Further data cleaning steps were conducted after translation, namely removal of conventions (# and @), and lower-casing. We also removed stop-words, except negations, removal of which might harm the overall classification accuracy.

Tonality off-the-shelf dictionaries

We investigated the performance of five off-the-shelf dictionaries for sentiment classification. Typically, simple ratios (e.g. share of words with positive or negative emotions) or count scores (e.g. number of words) are computed.

First, we wanted to assess the performance of the dictionary included in the LIWC (version 2007), which is widely used in social science and psychology. Second, we analysed the performance of the NRC Emotion Lexicon (Mohammad and Turney 2013), a crowdsourced dictionary of sentiments which includes both unigrams and bigrams. Third, we included AFINN (Nielsen 2011), a dictionary dedicated to analysing tweets which emphasize acronyms and other expressions used in microblogging, such as emotional reactions (‘LOL’ and ‘WTF’). Fourth, we analysed Lexicoder, a dictionary designed specifically for political text (Young and Soroka 2012). Lastly, we included Hu Liu’s dictionary (Hu and Liu 2004), which was created on the basis of a set of seed
adjectives (‘good’ and ‘bad’) and semantically expanded by applying the dictionary with synonymy and antonym relations provided by WordNet. The choice of using these dictionaries was based on the domain specificity of our corpus of tweets and on the language, as most lexical resources have been developed for English.

**Custom dictionary for tonality and democracy dimension**

The study on tonality about democracy (i.e. positive/neutral/negative connotations) was also conducted to extract several democracy dimensions. We focused on 11 dimensions theoretically based on the previous work of Kriesi et al. (2013) and Ferrín (2018) for the rounds six and ten of the European Social Survey (ESS).

These dimensions were adapted to the context of Twitter discussions. We grouped vertical and horizontal accountability, we differentiated voice and institutional participation, we added a category of sovereignty, and we did not differentiate between social and political equality. Compared to the survey items, our coding scheme differed in three further ways. First, we had to group some dimensions. The categories ‘freedom’ and ‘rules’ form two different survey dimensions, whereas they are grouped into one in our coding. Furthermore, the categories ‘voice’ and ‘sovereignty’ are absent from the survey items, but we introduced them in our coding scheme as they represented important dimensions in social media texts. Moreover, we did not differentiate between horizontal and vertical accountability when coding the tweets, as both dimensions were hard to disentangle in social media texts (see our final coding scheme in Table 1). Democracy could be framed along a number of other dimensions. We focus on these ones for the sake of our methodological study. However, the procedure outlined below can be applied to any dimension or concept, and is not necessarily limited to democracy.

A list of initial seed words associated with these dimensions were thus included in our democracy dimensions custom dictionary. We used WordNet to look for other words that were obvious synonyms and antonyms of the words included in the dictionary. Then, we followed the strategy elaborated by Amsler (2020), consisting in embedding the words from our dictionary into the overall corpus (not only a sample of manually annotated data) using word2vec (Mikolov et al. 2013) as implemented in the gensim library from Python. The general idea of word embedding is to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence (e.g. man is to king as women is to X). We trained our own model based on the collected tweets and we used Amsler’s LexExpander and LexEmbedder tools to further expand our dictionary.

For each entry in our dictionary, we determine whether they appeared in the sample of manually coded tweets and calculated a simple ratio of how many times they were coded as negative (\(w_n\)) or positive (\(w_p\)). This gives us the following equation: \(\sum w_p - \sum w_n\). The entry is labelled as positive if the overall score is above 0, and negative if below 0. A neutral category was added for entries that do not appear in the sample of annotated tweets or if the ratio of negative and positive tweets is equal.

Our final dictionary was composed of 11 democracy dimensions and 1776 words or expressions distributed among our dimensions. Technically, the matching of dictionaries was implemented using software liwcalike() from the quanteda package for R (Benoit et al. 2018).
| Title | Coding label | Keyword examples | Description |
|-------|--------------|------------------|-------------|
| Competition & fairness of electoral procedures | competition | election, campaign, candidate | Guarantees that elections are competitive, free, open, and fair (precondition for responsiveness). It also refers to the fact that the opposition is free to criticize the government. |
| Representation | representation | elected representatives, governing coalition, congressmen | Representation of citizens in the formal institutions, which translates in the fact that elected political actors act in citizens’ interest. The notion also refers to the crisis of democracy, but also to the type of representation (e.g. majority vs. proportional). |
| Self-organized participation of citizens | voice | social movement, strike, riot | Citizen’s self-organization into collective actions and reinventing political activism. |
| Participation of citizens through democratic procedures | institutional participation | direct democracy, participatory budget, e-voting | Citizens’ views are voiced in the political dialogue to be heard by political authorities (e.g. parliament, government, court) through institutional channels (e.g. plebiscite, popular initiative, referendum, petition, e-voting, etc.). |
| Social fairness & social or political equality | fair | redistribution, social justice, welfare | Social equality is the elimination of social and economic differences (e.g. poverty, income inequality) that would stand in the way of political equality. Political equality ensures that all citizens have equal opportunities to participate politically and to access the law. |
| Efficiency | efficiency | laundering, embezzlement, corruption | Implies that the democratic system discourages political authorities’ corruption and mismanagement of resources, but also guarantees parliament and government independence (e.g. from lobbies, interest groups, mafia, etc.). |
| Vertical & horizontal accountability | accountability | transparency, blameworthiness, limited power | Vertical accountability refers to mechanisms through which the people control their representatives. It also implies that elected politicians account, be responsible, answer for their political decisions and give justifications form their policy choices. Horizontal accountability refers to the checks and balances on government power via the courts. |
| Responsiveness to citizens | responsiveness to citizens | answerability, solution oriented, public opinion | Refers to the capacity to form and implement policies that citizens want (also accounts for democratic technologies such as e-democracy, e-governance, digital-government, etc.). |
| Responsiveness to other stakeholders | responsiveness to other stakeholders | agreement, convention, freedom of movement | Responsiveness can also be understood in terms of answerability to other governments, supranational bodies (e.g. EU, UN, etc.) or agreement (e.g. Schengen-Dublin). |
| Freedom & Rule of law | rules | court, freedom of speech, impartiality | Refers to the whole set of rights and liberties available to citizens (e.g. minority rights). The courts treat everyone the same, thus implying equality before the law. Gives information about what citizens think are essential characteristics of democracy (e.g. freedom of expression, freedom of the press, reliability of the press, etc.). |
| Sovereignism & nationalism | sovereignty | self-determination, nationalization, autonomy | Accounts for the competing views of liberal democracy, such as nationalism and sovereignism (defending a nation’s political autonomy, preserving cultural identity, and shielding the domestic economy), which are often conflated with populism or Euroscepticism. |
Manual coding of a sample of tweets

We coded 1426 randomly extracted tweets from our full dataset. The coding process was carried out by a single expert coder and was composed of two main tasks. First, each tweet was coded according to its tonality under three categories (positive, negative, and neutral). Second, each tweet was assigned to its main democracy dimension based on the coding scheme shown in Table 1. Regarding tonality, the sample of annotated tweets entails slightly more negative tweets (n = 729) than positive tweets (n = 646). With respect to democracy dimensions, annotated tweets tend to be skewed towards three categories, namely ‘institutional participation’ (n = 263), ‘representation’ (n = 224), and ‘rules’ (n = 217), followed by ‘accountability’ concerns (n = 161), ‘responsiveness to citizens’ (n = 154), ‘competition’ (n = 131), ‘voice’ (n = 78), ‘responsiveness to other stakeholders’ (n = 76), ‘fair’ (n = 60), ‘efficiency’ (n = 37), and ‘sovereignism’ (n = 23).

Chosen ML classification models.

We tested a set of ML models due to their conceptually different algorithmic approaches. The ROCCHIO (Rocchio 1971) classifier was implemented for both tonality and democracy dimensions tasks. ROCCHIO is a nearest centroid classifier applied to text classification. This method assigns to observations the label of the class of training samples which mean (centroid) is closest to the observation.

In addition, two different algorithms were used according to the specific task. Logistic regression, which is a classification algorithm used to solve binary classification problems, was used for tonality. For democracy dimensions we used SVM which are discriminative classifiers, fitting a margin-maximizing hyperplane between classes, and can be extended to non-linear problems of higher dimensionality using kernels that can accommodate any functional form.

In addition to supervised models, we also included an unsupervised BERT, which is a transformer for unsupervised language representation. Unlike word2vec, which generates a single word embedding representation for each word in the vocabulary, BERT considers the context for each occurrence of a given word. We pre-trained the BERT on the ensemble of collected tweets (n = 296,375). Therefore, unlike static resources, such as dictionaries, BERT calculates embeddings dynamically to represent the language data. We then conduct multinomial regression on the top of the BERT representation for predicting tonality and democracy frames.

Preparing the training and testing data sets

The annotated data are split into training and testing sets by a ratio of 80% (n = 1140) and 20% (n = 286). The performance of both ML and dictionary models is evaluated using the training/test approach via accuracy, precision, recall, and the F-score. Accuracy is the ratio of correctly predicted observation to the total observations. Precision indicates how many among the predicted positive are positive (true positive). In other words, it is the ratio of true positives over all those classified as positive. Recall is the ratio between true positive and all the actual positive units. The F-score combines precision and recall, and scores one under perfect precision and recall.
Results

**Tonality classification**

Focusing on tonality, Table 2 shows that an established dictionary, such as Lexicoder, is a good fit for the task at hand. However, the custom dictionary also performs very well, suggesting that we were able to include relevant features in the dictionary and to score them in a relevant direction. Nonetheless, dictionaries necessarily limit the amount of information that can be learned from the text. This is shown in Table 2, where BERT and ROCCHIO models outperform the dictionaries. ML, based on logistic regression, comes in second place for accuracy.

In contrast, other existing dictionaries (HuLiu, AFINN, and NRC) perform less well on the sample of annotated data. Furthermore, LIWC does not exceed random chance to predict tonality on our sample. The lower performance of some dictionaries (HuLiu, AFINN, LIWC07, NRC) compared to ML is notably due to the context of study. LIWC contains negative and positive words based on reviews that follow a stringent logic, contain little noise, and carry emotion-laden words. However, the HuLiu dictionary, which is constructed around adjectives to detect the emotional orientation of customer reviews, performs better. This can be since the HuLiu includes processing properties that can cope with misspellings and morphological variations. The low performance of existing dictionaries can be explained since, in our real-world application of tweets about democracy, the tonality of a text tends to be conveyed in less obvious ways. However, this does not necessarily invalidate their usefulness for predicting tonality in other circumstances. It nevertheless points to the necessity of testing ‘off-the-shelf’ dictionaries before we can derive conclusions from them. The number of missing matches in our custom dictionary are essentially due to very unusual writing styles and the use of multiple languages in the same tweet, which can lead to poor translations.

**Classification of the democracy dimensions**

Using a similar procedure to that for tonality, we compared dictionary and ML methods to classify tweets into the democracy dimensions. Because no off-the-shelf democracy dictionaries exist to detect our theoretical dimensions, for this task only our custom dictionary was used for the dictionary-based method.

Table 3 summarizes the accuracy of all democracy dimensions models. The average accuracy over all frame categories is as follows: 86% for the custom dictionary, 69% for BERT, 69% for ROCCHIO, and 66% for SVM. Our custom dictionary outperforms the other classification approaches for almost all democracy dimensions. BERT also performs generally better than ROCCHIO and SVM models.

**Table 2.** Classification results for predicting the tonality of tweets related to democracy.

| Precision | Recall | F-measure | Accuracy | Model   | Missing |
|-----------|--------|-----------|----------|---------|---------|
| 0.92      | 0.87   | 0.89      | 0.90     | BERT    | 0       |
| 0.86      | 0.89   | 0.88      | 0.88     | LR      | 0       |
| 0.85      | 0.78   | 0.81      | 0.83     | ROCCHIO | 0       |
| 0.95      | 0.65   | 0.77      | 0.81     | Lexicoder | 91     |
| 0.83      | 0.74   | 0.78      | 0.79     | CustomDict | 35    |
| 0.84      | 0.61   | 0.70      | 0.74     | HuLiu   | 199     |
| 0.83      | 0.58   | 0.69      | 0.72     | AFINN   | 177     |
| 0.85      | 0.55   | 0.67      | 0.71     | LIWC07  | 790     |
| 0.91      | 0.27   | 0.42      | 0.62     | NRC     | 24      |
There are large differences in maximum accuracies between the different classification methods with respect to the ‘efficiency’ and ‘sovereignty’ categories, which entail very few annotated data and impede more sophisticated classification methods to ‘learn’ from the data. The dictionary-based classifiers do not need to learn custom classes automatically from training data. Instead, they require expert-crafted dictionaries for such purposes and are likely to perform well if the scope of the included features has a good coverage of the research dimensions.

**Factors modelling the probability of a true value**

The previous descriptive findings and statistical tests suggest certain plausible explanations for these differences. However, the different potential explanations for the
observed accuracy differences cannot be disentangled. To further investigate which factors were possible drivers of performance, we ran a series of logistic regression models across the test dataset with accuracy of predicting human coding (correct vs. incorrect) as the dependent variable.

Text level factors, such as the number of words in the tweets, the number of hashtags (i.e. over or under five hashtags per tweet), the presence of mentions in the tweets, the original language of the tweets, the annotated sentiment, as well as the mention of any ambiguity encountered during the manual annotation process and whether the tweets contained an interrogation mark, were all inserted in the model as predictors.

With respect to the detection of tonality, we focus on the five best models (see columns 2–6 in Table 4). The regression results are displayed in Table 4 and show that tweets manually labelled as negative are more likely to be wrongly predicted by our custom dictionary and Lexicoder models in comparison to positive labelled tweets, although the reverse trend applies to ROCCHIO. Looking at the original language of tweets, French tweets were also more likely to be wrongly classified by our custom dictionary than German tweets. BERT and ROCCHIO also struggled to correctly assign tonality for tweets containing an interrogation mark.

With respect to the detection of democracy dimensions, we focus on the four best models (see columns 7–10 in Table 4). The regression results are displayed in Table 4 and show that the original language of tweets plays an essential role for the custom dictionary: French, and tweets in other languages, were more likely to be correctly classified by our custom dictionary than German tweets. Furthermore, the mention of an ambiguity during the manual annotation process leads tweets to be wrongly classified in every model. In addition, when looking at the tonality of the tweet, neutral tweets are more likely to be wrongly predicted by our custom dictionary and ROCCHIO in comparison to positive labelled tweets.

The results presented in Table 3 also point to the possible complementarity of the methods for text classification. Indeed, triangulating the prediction from the best dictionary and ML models could compensate the respective weaknesses of each approach. For instance, the methodology used in this article to induce and expand the custom dictionary can lead to the inclusion of terms that refer to several (albeit similar) concepts. However, ML models, especially BERT, are better at capturing single dense concepts that required more contextual understanding. Dictionaries could thus better identify the tweets containing rare (or complex) terms that can be matched with more than one translation, while ML models are better at integrating the contextual information.

Comparing surveyed attitudes with the prevalence of social media discussions

So far, we have applied multiple classification methods and tested for minimum levels of accuracy to determine the best classifier. In this section, we aim to compare surveyed attitudes from respondents to round six of the ESS on items covering the democracy dimensions that inspired the coding scheme for the extracted social media messages.

Concerning the measurement of democracy dimensions prevalence from Twitter messages, we select the best dictionary – our custom dictionary – and the best classifier – BERT –, and we apply them to the entire corpus to derive the prevalence of democracy
Table 4. Logistic regression on method performance to predict tonality (left) and democracy dimensions (right) as a function text and coding process characteristics.

| Characteristics                                      | Tonality                  | Democracy dimensions         |
|------------------------------------------------------|---------------------------|------------------------------|
|                                                      | CustomDict | Lexicoder | BERT | ROCCHIO | LR | CustomDict | BERT | ROCCHIO | SVM |
| (Intercept)                                          | 1.36 *      | 2.41 ***  | 1.31 | 1.35    | 0.62 | 1.32 *      | 1.16 * | 1.47 **  | 0.93 |
|                                                      | (0.65)      | (0.71)    | (0.87) | (0.73)  | (0.81)| (0.63)      | (0.54) | (0.55)  | (0.54) |
| Number of words in tweet                             | −0.00       | 0.00      | 0.03  | 0.01    | 0.06 *| −0.01       | −0.02  | −0.02   | −0.03 |
|                                                      | (0.02)      | (0.02)    | (0.03) | (0.02)  | (0.02)| (0.02)      | (0.02) | (0.02)  | (0.02) |
| Tweet contains hashtags (ref: <5)                    | −0.16       | 0.51      | 0.35  | −0.38   | −0.36| 0.57        | −0.20  | −0.16   | −0.25 |
|                                                      | (0.41)      | (0.48)    | (0.65) | (0.43)  | (0.50)| (0.44)      | (0.35) | (0.35)  | (0.35) |
| Tweet contains mention (ref: no)                      | −0.06       | 0.10      | 0.84  | 0.72    | −0.07| −0.36       | 0.04   | 0.07    | 0.12 |
|                                                      | (0.30)      | (0.33)    | (0.48) | (0.37)  | (0.40)| (0.31)      | (0.27) | (0.27)  | (0.27) |
| Original language (ref: German)                       |             |           |       |         |     |             |       |         |     |
| French                                               | 0.69 *      | −0.04     | −0.69 | −0.52   | −0.43| 0.75 *      | 0.16   | −0.15   | 0.39 |
|                                                      | (0.35)      | (0.37)    | (0.52) | (0.39)  | (0.45)| (0.34)      | (0.30) | (0.30)  | (0.30) |
| Other                                                | 0.45        | −0.29     | −0.86 | −0.39   | 0.15 | 1.22 **     | −0.10  | −0.26   | 0.20 |
|                                                      | (0.38)      | (0.39)    | (0.52) | (0.43)  | (0.53)| (0.40)      | (0.31) | (0.31)  | (0.31) |
| Annotated sentiment (ref: positive)                   | −0.68 *     | −2.21 *** | 0.60  | 0.73 *  | −0.39| −0.49       | −0.09  | −0.46   | −0.42 |
|                                                      | (0.29)      | (0.37)    | (0.42) | (0.33)  | (0.39)| (0.29)      | (0.25) | (0.25)  | (0.25) |
| Ambiguity in coding the tweet (ref: no)               | 0.33        | 0.09      | 0.07  | −0.76   | 0.52 | −1.35 **    | −1.11 **| −1.28 ** | −1.06 *|
|                                                      | (0.47)      | (0.50)    | (0.68) | (0.45)  | (0.78)| (0.42)      | (0.42) | (0.43)  | (0.44) |
| Tweet contains a question mark (ref: no)              | −0.46       | −0.18     | −0.97 *| −0.90 * | −0.36| 0.65        | 0.20   | 0.55    | 0.45 |
|                                                      | (0.34)      | (0.37)    | (0.48) | (0.39)  | (0.48)| (0.53)      | (0.40) | (0.42)  | (0.40) |
| AIC                                                   | 329.02      | 291.52    | 193.01| 269.39  | 214.09| 320.60      | 398.44 | 390.10  | 400.43|
| BIC                                                   | 361.98      | 324.49    | 225.98| 302.35  | 247.06| 353.56      | 431.40 | 423.06  | 433.40|
| Log Likelihood                                        | −155.51     | −136.76   | −87.51| −125.69 | −98.05| −151.30     | −190.22| −186.05 | −191.21|
| Deviance                                              | 311.02      | 273.52    | 175.01| 251.39  | 196.09| 302.60      | 380.44 | 372.10  | 382.43|
| Num. obs.                                             | 288         | 288       | 288   | 288     | 288  | 288         | 288    | 288     | 288  |

Note: Significativity levels are coded as follows: ‘***’ p < .001, ‘**’ p < .01, ‘*’ p < .05.
dimensions from unclassified tweets (those not labelled). We selected only those tweets that were equally classified by both methods (43,859 tweets in total). The sovereignty dimension is not represented among these tweets as the two classification methods did not reach agreement on this dimension. This can be explained by the fact that the sovereignty dimension was hardly represented in the annotated corpus and could only be identified by our custom dictionary. Conversely, the dimension accounting for rules of law and freedom related concerns is more prevalent in the corpus than any other dimension.

Concerning the measurement of the prevalence of democracy dimensions among survey respondents, we first had to choose between two possible measurements. The ESS distinguishes between people’s beliefs and expectations about what a democracy should be, asking whether different aspects are important for the workings of democracy, and people’s evaluations of their own democracies with respect to these aspects. We chose to concentrate on respondents’ rated importance of the democracy dimensions as it compares better conceptually to the estimated prevalence of democracy dimensions in social media. We organized the survey items related to democracy into the corresponding democracy dimensions following the template from Kriesi et al. (2013). However, we remove items that are negative formulations of positive items and items that are at odds with the functioning of democracy in Switzerland. The response scale to rate each aspect range from 1 ‘not important at all’ to 11 ‘extremely important’. We selected only respondents who positioned themselves on all survey items (N = 686). To be able to compare the survey results with social media findings, we only considered answers above seven (excluded) when rating the importance of each dimension for the workings of democracy. We took the mean if more than one item represented the dimension.

To be able to directly compare survey and social media prevalence on the common democracy dimensions, we do a correlation analysis (see Figure 1). We computed Spearman rank correlation between the number of tweets classified in each dimension and the number of answers with a score higher than 7. The correlation coefficient is 0.58, confirming a positive correlation between the opinion surveyed in the ESS and those spontaneously expressed on Twitter. It also shows that a moderate proportion of variance differs between the scores extracted from the tweets and those in the ESS, supporting the argument for a complementarity of the two methods. Social media users and survey respondents emphasize the category related to the rule of law. Political accountability and competition are also perceived as important for the workings of democracy. On the opposite, responsiveness to other stakeholders (mainly the European Union) is considered least important.

**Discussion and concluding remarks**

*_Where to use dictionaries and where to use ML?_*

The goal and main contribution of this manuscript is to provide easily implementable recommendations for increasing estimation accuracy under non-optimal conditions. With this goal in mind, we discuss how to run and compare different methods, in order to make informed and well-suited decisions. In this article, we develop an empirical application relating to tweets about democracy, from which we detect their tonality (positive versus negative sentiment) and frames (substantive democracy dimensions). We have
relied on two main approaches for conducting both sentiment and framing analysis, namely supervised ML, and dictionary-based frameworks. We also employ a third approach which consists in using unsupervised classification methods relying on word embeddings to classify text by word vector similarity. These approaches allow researchers to code vast amounts of text that would not be possible with manual coding, and each presents unique advantages but also challenges.

Our main goal was to assess how well these approaches perform on a small and skewed sample of annotated data, and how off-the-shelf dictionaries compare to ‘custom’ dictionaries to accomplish these classification tasks. Our objective was not to build ‘state-of-the-art’ classifiers with optimal performance, but to provide a procedure with which to understand how methods utilizing minimal training data compare against off-the-shelf dictionaries.

In our study, the dictionary methods lagged behind BERT and ROCCHIO models for tonality classification, but performed better when the task was more complex, such as classifying tweets according to several democracy dimensions. This can be explained partially by the over-representation of negative tonality across all democracy dimensions (except for the institutional participation) in our sample of annotated tweets. Indeed, the fact that Twitter users tend to post more on aspects of democracy which are perceived could have ‘nudged’ classifiers in a negative direction, thereby enabling ML methods to classify negative tweets more correctly than dictionary-based approaches. This finding is confirmed by the regression analysis on the prediction accuracy, which shows that ML classify tonality more easily than dictionaries.

On the contrary, our custom dictionary performed well for detecting democracy dimensions as it contains words that contribute significantly to these dimensions.

Figure 1. Relation between the importance of democracy dimensions for democracy in the ESS (x-axis) and the number of tweets related to each dimension (y-axis).
The dictionary-based approach could thus serve as a useful complement to ML classifiers, which rely on every word feature contained in the training dataset fed in. However, if the reliance of ML models on the entire training dataset has the advantages of covering many more word features than the ones present in the dictionary, it may also be problematic when there is a bias toward most prevalent categories (skewness). This result should not be interpreted as indicating that dictionary methods are better than ML in absolute terms. On the contrary, we think that BERT models provide the most appropriate solution to date. In our analysis, BERT was indeed the most reliable among the ML models, though it performed less well than the custom dictionary method for very specific categories that were under-represented in our sample (e.g. ‘sovereignism’). In cases of large training data and/or balanced categories, the ML methods might have outperformed against the dictionary ones. However, this is a context-dependent aspect which cannot be generalized, but needs to be investigated each time. Importantly, the fact that ML approaches tend to be error-prone for classes with fewer examples should not mean that lower performance is acceptable as long as the error is distributed in an unbalanced way. A possible solution would be to use the dictionary to over-sample the under-represented categories in the training set before ML models are employed.

**What best practices can we derive from our case study?**

Our results showed the usefulness of testing different approaches to carrying out a classification task (in our case, we took the best dictionary and the best unsupervised classifier), as it enables us to assess which classification method is more suited to the classification task at hand. Instead of blindly relying on preconstructed methods, a train-and-test framework can be easily set to empirically judge the accuracy and reliability of each method. Even the use of a small, annotated data set might be useful for understanding which among different methods of estimation or different dictionaries could be the most appropriate for the goal of the study. If larger training datasets are preferable, good proxies can lead to reliable results, even under poor research conditions.

We also showed a procedure of building custom dictionary which complements embeddings techniques and human judgement. To construct the dictionary, we went through several steps by, first, departing from a few seed words derived from the literature and from mainstream media to which we added salient terms and hashtags from the sample of annotated tweets and, second, we trained a word vector model (see Amsler’s (2020) LexEmbedder and LexExpander) based on the entire collection of tweets to add eligible word candidates to our dictionaries. This process involved a continuing translation between human validation and computational suggestions. We believe this method might be of particular interest for social and political scientists because it creates a loop between data-driven findings and theory-driven choices. On the one hand, it allows to build more exhaustive dictionaries, and on the other gives the researchers a control over the concepts they want to measure, relying less on black-box algorithms and estimating possible imbalances in the data.

The translation process of the tweets into English has a non-zero impact on the results, albeit it is difficult to really assess how. While we could have elaborated our custom dictionaries in the original languages of our corpus and translated the off-
the-shelf dictionaries term-wise to apply these resources in the original languages of the tweets, there is an opportunity to seek a minimal transfer cost. Most notably, there is a non-negligible issue arising from the multitude of possible translations for single terms in each of the chosen languages (Vicente and Saralegi 2016). Concerning the application of ML models, a possible solution would have been to align the text in each of the original languages contained in our data and to conduct separate models for single languages. Other solutions might include relying on transfer learning or cross-lingual embeddings, but this would require a non-negligible amount of training data.

Before we conclude, we would like to point to several aspects that are worth investigating in future case studies. The manual verification of the misclassified data by both methods enabled us to derive several common difficulties. For instance, both methods had trouble tackling specific linguistic patterns, such as double negations, thwarted expression, and negated expression. Future research should try to alleviate these difficulties using tagging and parsing. Furthermore, some tweets contained a dose of irony which complicated the classification task. Finally, we advise researchers to conserve as much as possible the language of the original tweets, as translations can introduce false synonyms and these translations may not always be of good quality.

Notes
1. Data retrieved from Web of Science.
2. Similar projects have used a dictionary-based approach to detect frames. We mention a sample. For instance, Maerz (2019) studied democratic systems relying on the analysis of the language of authoritarian leaders. Laver, Benoit, and Garry (2003) developed a dictionary of policy positions, where words have been selected semantically, based on how they relate to specific content categories as well as to a specific political party. Kraft (2018) explored whether and how individuals evoke moral considerations when discussing their political beliefs (e.g., harm, fairness, ingroup, authority, purity, and morality).
3. Swiss German speaking quality newspaper (Neue Zürcher Zeitung), seven regional newspapers (Tagesanzeiger, Aargauer Zeitung, Basler Zeitung, Berner Zeitung, Neue Luzerner Zeitung, Die Südostschweiz, St. Galler Tagblatt), one free newspaper (20 Minuten), one tabloid press (Blick), and three Sunday press (Sonntagsblick, NZZ am Sonntag, Sonntagszeitung).
4. One Swiss French speaking quality newspaper (Le Temps), seven regional newspapers (Tribune de Genève, 24 heures, La Liberté, Le Nouvelliste, Le Courrier, L’Express, L’Impartial), one free newspaper (20 Minutes), one tabloid press (Le Matin), and one Sunday press (Le Matin dimanche).
5. We also noted that depending on whether the grammatical and syntactic rules are very different from the resulting language, the translation can be far from the original word ordering, sometimes losing the meaning of the text. Furthermore, while some words can be suitable for direct translation into a single term, other words are complex to translate. In these cases, the automatic translator might be incentivised to translate with the most commonly used words.
6. WordNet (Miller 1995) is a large lexical database of English. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. See more here: https://wordnet.princeton.edu/
7. If no previous concepts are available and the dictionary must be elaborated from scratch, software such as WordNet (Miller 1995) can be used to collect main concepts.
8. WordNet (Miller 1995) is not available in many languages in such comprehensive versions as for English.
9. The traditional approach towards opinion mining or text classification in general (also called Bag-Of-Words approach) considers a sentence (or a document) as Bag containing words. It considers the words and their frequency of occurrence in the document, disregarding semantic relationship in the sentences (albeit the Bag can grow by searching for synonyms and antonyms).

10. A tweet was coded as neutral only if it contains an announcement of an event or a brief journalistic review. In total, we coded 50 tweets as neutral.

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

ORCID
Maud Reveilhac http://orcid.org/0000-0001-9769-6830
Davide Morselli http://orcid.org/0000-0003-3231-7769

References
Albaugh, Q., J. Sevenans, S. Soroka, and P. J. Loewen. 2013. “The Automated Coding of Policy Agendas: A Dictionary-based Approach”. In Proceedings of the sixth Annual Comparative Agendas Conference, 1-27.

Amsler, M. 2020. Using Lexical-Semantic Concepts for Fine-Grained Classification in the Embedding Space. Doctoral dissertation. University of Zurich, Faculty of Arts.

Amsler, M., B. Wüest, and G. Schneider. 2016. “Legitimacy of New Forms of Governance in Public Discourse - An Automated Media Content Analysis Approach Driven By Techniques of Computational Linguistics”. In Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText), 1–7.

Barberá, P., A. Boydstun, S. Linn, R. McMahon, and J. Nagler. 2021. “Automated Text Classification of News Articles: A Practical Guide.” Political Analysis 29 (1): 19–42. doi:10.1017/pan.2020.8

Benoit, K., K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo. 2018. “quanteda: An R Package for the Quantitative Analysis of Textual Data.” Journal of Open Source Software 3 (30): 1–4. doi:10.21105/joss.00774

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. “Latent Dirichlet Allocation.” Journal of Machine Learning Research 3: 993–1022.

Card, D., J. H. Gross, A. Boydstun, and N. A. Smith. 2016. “Analyzing Framing Through the Casts of Characters in the News”. In Proceedings of the conference on empirical methods in natural language processing, 1410-1420.

Cortes, C., L. D. Jackel, S. A. Solla, V. Vapnik, and J. S. Denker. 1994. “Learning Curves: Asymptotic Values and Rate of Convergence”. In Proceedings of the sixth International conference on neural information processing system, 327-334.

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2018. “Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding”. arXiv preprint arXiv:1810.04805.

DiMaggio, P., M. Nag, and D. Blei. 2013. “Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of US Government Arts Funding.” Poetics 41 (6): 570–606. doi:10.1016/j.poetic.2013.08.004

Ferrin, M. 2018. European Social Survey Round 10 Module Design Teams (QDT) Stage 2 Application. London: Centre for Comparative Social Surveys, City University London. http://www.europeansocialsurvey.org/docs/round10/questionnaire/ESS10_ferrin_proposal.pdf.

Fishman, R. M. 2016. “Rethinking Dimensions of Democracy for Empirical Analysis: Authenticity, Quality, Depth, and Consolidation.” Annual Review of Political Science 19: 289–309. doi:10.1146/annurev-polisci-042114-015910
García-Marín, J., and A. Calatrava. 2018. “The use of Supervised Learning Algorithms in Political Communication and Media Studies: Locating Frames in the Press.” Comunicación y Sociedad 31 (3): 175–188.

Gillard, F., T. Gessler, M. Kubli, and S. Müller. 2021. “Social Media and Political Agenda Setting.” Political Communication, 1–22. doi:10.1080/10584609.2021.1910390

Gilardi, F., Shipan, C. R., & Wüest, B. 2021. “Policy Diffusion: The Issue–definition Stage.” American Journal of Political Science 65 (1): 21–35. doi:10.1111/ajps.12521.

Gilardi, F., and B. Wüest. 2018. “Using Text-as-Data Methods in Comparative Policy Analysis”. In Handbook of Research Methods and Applications in Comparative Policy Analysis, edited by G. Peters, and G. Fontaine, 203–217. Northampton: Edward Elgar Publishing.

Goffman, E. 1974. Frame Analysis: An Essay on the Organization of Experience. New York, NY: Harper & Row.

Grimmer, J., and B. Stewart. 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

Haixiang, G., L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing. 2017. “Learning from Class- Imbalanced Data: Review of Methods and Applications.” Expert Systems with Applications 73: 220–239. doi:10.1016/j.eswa.2016.12.035

Hamilton, W. L., K. Clark, J. Leskovec, and D. Jurafsky. 2016. “Inducing Domain Specific Sentiment Lexicons from Unlabeled Corpora”. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 595-605.

Hartmann, J., J. Huppertz, C. Schamp, and M. Heitmann. 2019. “Comparing Automated Text Classification Methods.” International Journal of Research in Marketing 36 (1): 20–38. doi:10.1016/j.ijresmar.2018.09.009

Hu, M., and B. Liu. 2004. “Mining and Summarizing Customer Reviews”. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168-177.

Iliev, R., M. Dehghani, and E. Sagi. 2015. “Automated Text Analysis in Psychology: Methods, Applications, and Future Developments.” Language and Cognition 7 (2): 265–290. doi:10.1017/langcog.2014.30

Japec, L., F. Kreuter, M. Berg, P. Biemer, P. Decker, C. Lampe, J. Lane, C. O’Neil, and A. Usher. 2015. “Big Data in Survey Research: AAPOR Task Force Report.” Public Opinion Quarterly 79 (4): 839–880. doi:10.1093/poq/nfv039

Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov. 2017. “Bag of Tricks for Efficient Text Classification”. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2, 427-431.

Kraft, P. W. 2018. “Measuring Morality in Political Attitude Expression.” Journal of Politics 80 (3): 1028–1033. doi:10.1086/696862

Kriesi, H., L. Morlino, P. Magalhães, S. Alonso, and M. Ferrín. 2013. European Social Survey Round 6 Module on Europeans’ Understandings and Evaluations of Democracy – Final Module in Template. London: Centre for Comparative Social Surveys, City University London. https://www.europeansocialsurvey.org/docs/round6/questionnaire/ESS6_final_understandings_and_evaluation_of_democracy_module_template.pdf.

Laver, M., K. Benoit, and J. Garry. 2003. “Extracting Policy Positions from Political Texts Using Words as Data.” American Political Science Review 97 (2): 311–331.

Loughran, T., and B. McDonald. 2011. “When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks.” The Journal of Finance 66 (1): 35–65. doi:10.1111/j.1540-6261.2010.01625.x

Maerz, S. F. 2019. “Simulating Pluralism: The Language of Democracy in Hegemonic Authoritarianism.” Political Research Exchange 1 (1): 1–23. doi:10.1080/2474736X.2019.1605834

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. “Distributed Representations of Words and Phrases and Their Compositionality”, arXiv preprint arXiv:1310.4546.

Miller, G. A. 1995. “WordNet: A Lexical Database for English.” Communications of the ACM 38 (11): 39–41. doi:10.1145/219717.219748

Mohammad, S. M., and P. D. Turney. 2013. “Nrc emotion lexicon”. National Research Council, Canada, 2. http://www.saifmohammad.com/WebDocs/NRCemotionlexicon.pdf.
Annex 1: List of search-queries to collect the tweets

| Search-queries | French | German | Italian | Hashtags |
|---------------|--------|--------|---------|----------|
| **From press review and theoretical inputs** | | | | |
| 1 | democratie | demokratie | democrazia | |
| 2 | democratique OR democratiques | demokratisch OR demokratischen | democratico OR democratica | |
| 3 | anti-democratique OR antidemocratique | undemokratische OR undemokratischen | anti-democratico OR antidemocratica | |
| 4 | (Continued) | | | |
| Search-queries | French | German | Italian | Hashtags |
|----------------|--------|--------|---------|----------|
| 5              | democratie directe | direkte demokratie | democracia diretta |          |
| 6              | politique | politik | politico OR politica |          |
| 7              | politicien OR politiciens | politiker OR politikern | politici OR politiche |          |
| 8              | politicienne OR politiciennes | politikerin OR politikerinnen | deputato OR deputata OR deputati OR deputate parlamentari |          |
| 9              | politique | politik | politico OR politica |          |
| 10             | votants | wähler | elettori |          |
| 11             | citoyennete | staatsbürgerschaft | cittadinanza |          |
| 12             | citoyens | staatsbürger | cittadini |          |
| 13             | citoyen | stimmbürger | cittadino |          |
| 14             | people | volk | popolo |          |
| 15             | elite OR elite | eliten | elite |          |
| 16             | elu OR elus OR elue OR elues | volksvertreter OR volksvertretern | rappresentanti |          |
| 17             | conseiller OR conseillers OR conseillère OR conseillères | abgeordnete OR abgeordneten | rappresentante |          |
| 18             | parlementaire OR parlementaires | regierung | governo |          |
| 19             | conseil federal | bundesrat | consiglio federale |          |
| 20             | parlement | parlament | parlamento |          |
| 21             | assemblee federale | bundesversammlung | assemblea federale |          |
| 22             | representation AND politique | vertretung |          |          |
| 23             | constitutionnel OR constitutionnelle | verfassung | costituzionale |          |
| 24             | federal OR federale | federal | federal |          |
| 25             | fédérales OR fédéraux | staat OR staatlisch | diretti popolari |          |
| 26             | droits populaires | initiativ | iniziativa |          |
| 27             | vote | abstimmung | voto |          |
| 28             | votations | abstimmungen | voti |          |
| 29             | volk | volksbeteiligung | voti populari |          |
| 30             | scrutine | volksentscheidung |          |          |
| 31             | votations populaires | volksbefragung |          |          |
| 32             | elections | wahlen | elezioni |          |
| 33             | campagne AND electorale | wahlkampagne | campagna AND elettorale |          |
| 34             | bataille AND electorale | wahlkampf | concorrenza AND elettorale |          |
| 35             | referendum | referendum | referendum |          |
### Search-queries

|   | French | German | Italian | Hashtags |
|---|--------|--------|---------|----------|
| 53 | referendaires | referendaires | referendaires | referendaires |
| 54 | décision populaire | stimmvolk | stimmvolk | decision populaire |
| 55 | institutionnel OR institutionnelle | institutionel OR institutionelle | istituzionale OR istituzionale | institutionnel OR institutionnelle |
| 56 | vote OR voter | stimmen | stimmen | voitare |
| 57 | contre-projet | gegenvorschlag | partei OR parteien | controprogetto |
| 58 | parti politique OR partis politiques | partei OR parteien | parti OR parteien | partito OR partiti |
| 59 | le parti OR les partis | rechte AND partei OR parteien OR politik | sinistra AND partito OR partiti OR politica | destra AND partito OR partiti OR politica |
| 60 | droite AND parti OR partis OR politique | linke AND partei OR parteien OR politik | centro AND partito OR partiti OR politica | estrema AND destra OR sinistra |
| 61 | gauche AND parti OR partis OR politique | rechtsextrem OR linkeextrem | politische rechte OR politischen rechten | diritti politici |
| 62 | extrême OR extreme AND droite OR gauche droits politiques | koalition OR coalition | koalition OR coalition | coalizione |
| 63 | coalition | soverain | soverain | sovranità |
| 64 | souverain | soverän OR souveränen | soverain | sovrano |

From hashtags review

1 #abst18
2 #abst19
3 #Abstimmungssonntag
4 #AccordCadre
5 #Bundeshaus
6 #Bundesrat
7 #Cantonal19
8 #cantonali2019
9 #chwote
10 #ConseilFédéral
11 #ConsiglioStato
12 #ConsiglioFederale
13 #contreprojet
14 #controprogetto
15 #deardemocracy
16 #democrazia
17 #démocratieparticipative
18 #democrazia
19 #democraziadiretta
20 #Demokratie
21 #DirekteDemokratie
22 #eDemocracy
23 #ef19
24 #ef2019
25 #eGov
26 #eGovernment
27 #elezioni2019
28 #evoting
29 #eVotingMoratorium

(Continued)
Continued.

| Search-queries | French | German | Italian | Hashtags |
|----------------|--------|--------|---------|----------|
| 30             |        |        | #facciamolo |
| 31             |        |        | #gegenvorschlag |
| 32             |        |        | #granconsiglio |
| 33             |        |        | #iniziativa |
| 34             |        |        | #Lobby |
| 35             |        |        | #Lobbying |
| 36             |        |        | #nationalratswahlen |
| 37             |        |        | #NeinzumEURahmenabkommen |
| 38             |        |        | #NeinzurEU |
| 39             |        |        | #NoEvoting |
| 40             |        |        | #nrw19 |
| 41             |        |        | #parlCH |
| 42             |        |        | #parldigi |
| 43             |        |        | #participation |
| 44             |        |        | #politica |
| 45             |        |        | #politik |
| 46             |        |        | #politique |
| 47             |        |        | #Rahmenabkommen |
| 48             |        |        | #Rahmenvertrag |
| 49             |        |        | #Rechtsstaat |
| 50             |        |        | #referendum |
| 51             |        |        | #societé civile |
| 52             |        |        | #souveraineté |
| 53             |        |        | #Souveränität |
| 54             |        |        | #Sovranità |
| 55             |        |        | #SwissEUrelations |
| 56             |        |        | #swissregulations |
| 57             |        |        | #transparence |
| 58             |        |        | #transparenz |
| 59             |        |        | #Versicherungslobby |
| 60             |        |        | #verwaltung |
| 61             |        |        | #Volksabstimmungen |
| 62             |        |        | #Volksinitiativen |
| 63             |        |        | #vorwärts |
| 64             |        |        | #vot18 |
| 65             |        |        | #vot19 |
| 66             |        |        | #votations |
| 67             |        |        | #votazioni |
| 68             |        |        | #wahlCh19 |
| 69             |        |        | #wahlen19 |
| 70             |        |        | #wahlen2019 |
| 71             |        |        | #wahlenCH19 |