Leveraging Researcher Domain Expertise to Annotate Concepts within Imbalanced Data

Dror K. Markus* [a], Guy Mor-Lan* [a], Tamir Sheafer [a,b] and Shaul R. Shenhav [a]

[a]Department of Political Science, The Hebrew University, Jerusalem, Israel; [b]Department of Communication and Journalism, The Hebrew University, Jerusalem, Israel

* Both authors contributed equally to this work

Corresponding Author:

Dror Markus, dror.markus@mail.huji.ac.il, (972) 54-3369228

Leveraging Researcher Domain Expertise to Annotate Concepts within Imbalanced Data

As more computational communication researchers turn to supervised machine learning methods for text classification, we note the challenge in implementing such techniques within an imbalance dataset. Such issues are critical in our domain, where, in many cases, researchers attempt to identify and study theoretically interesting categories that can be rare in a target corpus. Specifically, imbalanced distributions, with a skewed distribution of texts among the categories, can lead to a lengthy and expensive annotation stage, forcing practitioners to sample and label large numbers of texts to train a classification model. In this paper, we provide an overview of the issue, and describe existing strategies for mitigating such challenges. Noting the pitfalls of previous solutions, we then provide a semi-supervised method that complements researcher domain expertise with a systematic, unsupervised exploration of the latent semantic space to overcome such limitations. Utilizing simulations to systematically evaluate our method and compare it to existing approaches, we show that our procedure offers significant advantages in terms of efficiency and accuracy in many classification tasks.

Introduction

Computational text analysis is becoming increasingly common in social science research. Most predominant are *unsupervised* text analysis algorithms that learn from textual corpora without any a priori annotations or guidance, seeking to uncover inherent patterns or inductively identify structure (Chatsiou and Mikhaylov, 2020; Grimmer and Stewart, 2013; Miller, Linder, and Mebane, 2020). Especially common are several methods for unsupervised topic modelling – which inductively identify a set of thematic topics in the corpus of interest. Unsupervised methods have proven to be relatively cheap, yet powerful and well-suited for exploratory or inductive research questions.

However, in recent years, we have seen a growing appreciation for *supervised* machine learning methods (e.g., Dobbrik et al., 2021; Gilardi et al., 2021; Guess et al., 2018; Merkley and Stecula, 2021; Pilny et al., 2019)[1]. In these methods, the human researcher compiles a training set of labelled samples for an algorithm to learn from, before developing a predictive model that can then be used to classify new, unlabeled instances (Chatsiou and Mikhaylov, 2020; Loftis and Mortensen, 2020). Supervised learning offers several advantages for communication researchers, most prominently, a deductive research approach providing an alternative to the inductive research strategy most often used with unsupervised machine learning methods. In this approach, researchers can emulate the procedures of traditional content analysis: developing a theory-based hypothesis, operationalizing variables of interest, and analyzing the target corpus with such categories *(*categories of interest can be referred to in the machine learning literature as *classes)* in mind (Bauer, 2000; Krippendorf, 2018). This greater control of the research design allows researchers to focus on a wide range of theoretical categories, including those corresponding directly to their research question (Fogel-Dror, Shenhav, and Sheafer, 2018; Miller, Linder, and Mebane, 2020; Mor-Lan, 2019).

While the structure of supervised learning seems to fit traditional communication research, the implementation of such procedures involves various costs and challenges. While unsupervised algorithms can be utilized on any new, unclassified corpus or dataset, in supervised learning, the researchers must prepare a collection of labeled samples (a training set) that represents and encompasses each target category. This training set is then used to train the

[1] While unsupervised learning remains more prevalent in Communication studies than supervised methods, we can observe a noteworthy increase in the latter's utilization. In 2015, only one article was published which included the phrase "supervised machine learning" in its abstract. In 2016 that number rose to five, and by 2021 fifteen articles were published with that phrase (Source: https://www.webofscience.com/wos/woscc/analyze-results/7f5f3405-e1ec-4bb1-8e27-2c7148ce6fec-52779423).

supervised model, which is then utilized to classify the remaining majority of unlabelled documents. However, many theoretically important concepts in our field may be quite rare in our target corpora. Such category imbalance can complicate the sampling and collection of sample texts for a training set, as well as the ensuing learning process (Stoll et al., 2020, p. 121). For example, a researcher attempting to identify categories such as *populist rhetoric* (Dai and Kustov, 2022, p. 8) or expressions of *impoliteness* and *incivility* (Stoll et al., 2020) in a corpus consisting of general media coverage or a collection of social media posts, may find themselves having to sort through multiple rounds of random sampling before amassing enough documents to utilize as a training set. Additionally, the amorphousness and complexity of the theoretical concepts being studied may spell difficulties in the selection of fully-encompassing and informative samples.

With these issues in mind, we propose a method for annotation of an unlabeled corpus with the considerations of communication researchers in mind. We will first provide an overview of the supervised approach, noting the advantages of such methods in specific research projects. We then describe the issue of minority classes and imbalanced data, reviewing the various solutions offered in the machine learning literature. We proceed to discuss a number of limitations of such solutions, before presenting a new method for compiling and annotating training sets (building upon existing techniques and approaches), suited specifically to handle theoretical concepts in the communication field and other social sciences. Our method first utilizes researcher domain expertise to manually compile a preliminary 'nucleus' of representative documents for each target category. Then, utilizing unsupervised document embeddings to allow us to place the texts of our corpus onto a shared semantic vector space (which we describe in greater detail below), we provide a systematic process consisting of

multiple iterations of computerized concept expansion with researcher validation that allows us to grow our original sample set. Such unsupervised diversification adds depth and nuance to the training set for each category, while also helping to mitigate researcher biases in the nucleus compilation. These training sets can then form the basis of a text classification model. Finally, we will utilize several simulations to demonstrate the efficacy of our method for particular research applications.

Thus, we hope to provide an easy-to-use method for the creation of training sets – improving efficiency and accuracy when working with rare, theory-driven concepts. Additionally, this research joins previous endeavors to deepen the symbiosis of human and computer methods (e.g. Grimmer and Stewart, 2013; Merchant, 2021; Ophir, Walter, and Merchant, 2020). Such interactions are recognized as an important step in improving the utilization and development of computational methods. The process described here is such an example of how to 'inject' doses of domain expertise, and thus improve greatly the efficiency of machine learning classification (as we will demonstrate in simulations below).

The opportunities of Supervised Learning

Machine learning methods utilized in computational communication research can be broadly divided into two major approaches: unsupervised and supervised learning. Unsupervised algorithms learn from the data without any a priori guidance or labeling, seeking instead to uncover latent patterns or inductively identify structure (Chatsiou and Mikhalov, 2020; Grimmer and Stewart, 2013; Miller, Linger, and Mebane, 2020). This approach includes a number of useful methods – including topic modelling, anomaly detection, and community detection within

networks (e.g., Nicholls and Bright, 2019; Trilling and van Hoof, 2020; Stoltenberg, Maier, and Waldherr, 2019; Walter and Ophir, 2020).

Unsupervised learning provides a number of important advantages. Such methods have proven to be relatively cheap, yet powerful and well-suited for inductive research questions (Miller, Linger, and Mebane, 2020). Researchers can also use such techniques at the start of their research – exploring the data and determining how well their initial theoretical conceptions fit the data (Ophir et al, 2021). Additionally, there has been important work allowing the utilization of unsupervised models' results as research variables, including formalized validation procedures for unsupervised models (e.g., Ying, Montgomery, and Stewart, 2021; Walter and Ophir, 2020).

While unsupervised machine learning provides inductive research solutions, supervised learning offers a second type of approach. In supervised learning, the human researcher compiles a training set of labelled samples for an algorithm to learn from and produce a predictive model that can then be used to classify new, unlabeled instances (Chatsiou and Mikhaylov, 2020; Loftis and Mortensen, 2020). This method is steadily gaining attention and increased utilization by computational communication researchers (e.g., Dobbrik et al., 2021; Gilardi et al., 2021; Guess et al., 2018; Merkley and Stecula, 2021; Pilny et al., 2019).

Supervised methods offer several advantages and are better suited for certain types of research. First, supervised learning provides an alternative, deductive research strategy instead of the inductive approach commonly implemented with unsupervised machine learning methods, allowing researchers to target specific, theory-driven categories (Fogel-Dror, Shenhav, and Sheafer, 2018; Miller, Linder, and Mebane, 2020; Mor-Lan, 2019). In this way, this approach corresponds closely to the standard research paradigm in social science, posing theory-based

hypotheses, operationalizing the related variables, and using measurements to test the hypothesis (Bhattacherjee, 2012). Specifically, in regard to text analysis, supervised learning can be thought of as an extension of classic content analysis: researchers decide which variables they wish to measure, create coding schemes, and only then approach the target texts for analysis (Bauer, 2000; Krippendorf, 2018). In supervised computational content analysis, researchers first define and operationalize their categories of interest, feeding the learning algorithm samples that represent such concepts, in order to develop models trained specifically to seek such classes in the target, unlabeled corpus. This stands in contrast to unsupervised topic modeling which creates clusters based on latent, underlying structures in the data, not necessarily with the theory-derived concepts in mind (Ying, Montgomery, and Stewart, 2021).

Whereas unsupervised models are usually engineered to produce a specific kind of output (e.g., identify thematic clusters, identify anomalous observations), supervised methods afford researchers great flexibility in determining the unit of analysis (e.g. word, sentence, paragraph, book), the nature of classification scheme and the interrelation between its categories, as well the model's architecture. Thus, researchers can decide whether their classification scheme is multiclass (attempting to target more than one category), multilabel (allowing data instances to belong to more than one category) (e.g., Tsoumakas and Katakis, 2007; for additional design creativity – Dekel and Shamir, 2010). While unsupervised clustering typically hones in on clusters of thematic content, in supervised learning a researcher can choose to focus on semantics, style and other structural features of the text (or even including multiple dimensions within a single category). This offers the opportunity to target important theoretical concepts which can span multiple themes and textual characteristics. For example, researchers exploring

political ideologies might seek a classification that encompasses themes such as foreign policy, social values and economic principles, and style when delineating *conservativism* or *liberalism*.

A final advantage of supervised learning is the ability to measure the quality of the model in a straightforward way (Grimmer and Stewart, 2013). Supervised methods compare their performance against a gold standard benchmark, such as a test set consisting of samples set aside and labeled before the learning procedure. This allows for a straightforward measurement of accuracy – the success of the model in correctly predicting a human validated test set. Thus, model performance is compared directly with human annotators. While there has been much work on developing quality measures for unsupervised models as well, such metrics are generally descriptive statistics of the output of a model, that do not necessarily correlate with the quality according to human interpretation. Some examples include coherence and perplexity in topic models (see discussion in Ying, Montgomery, and Stewart, 2021), and modularity in community detection (see Stoltenberg, Maier, and Waldherr, 2019). These metrics provide a statistical description of the output yet are not a direct measure of how accurate a model is in identifying the specific variables or categories targeted by the researcher (see, for example, Lau, Newman, and Baldwin, 2014).

Such features point to important advantages in utilizing supervised learning in particular research circumstances. Researchers adopting this approach need to take note of a crucial stage of supervised machine learning - the collection and annotation of samples for the training data (Pustejovsky and Stubbs, 2013; Grimmer and Stewart, 2013). As Grimmer and Stewart explain, "no statistical model can repair a poorly constructed training set" (2013: 276). Such a set must be large and diverse enough to represent each target category and allow the utilized model to distinguish between each. However, human annotation of documents may be prohibitively costly

in time and effort. Thus, researchers might seek to find ways to reduce the number of documents to label, while still maintaining accuracy (see discussions in Miller, Linger, and Mebane, 2021; Loftis and Mortensen, 2020; Sebők and Kacsuk, 2021). Such parameters can be difficult to ascertain a priori – researchers do not necessarily know how many training samples are required,[2] nor which samples best represent the target concept.

This issue is compounded in cases of imbalanced datasets and rare categories. A basic approach to creating a training set is to randomly sample the data for instances for human labeling. However, such a strategy is not optimal in the case of imbalanced data, as it might take many multiple rounds of random sampling to collect enough samples for each of the classes (Miller, Linger and Mebane, 2020). Data imbalance is of significance to the machine learning community at large, with several approaches and solutions which we detail in the next section.

Working with Imbalanced Data

While machine learning algorithms assume a relatively balanced distribution of classes within a dataset – a similar number of instances for each category – the real world can be 'messier'. In many cases, the distribution of categories is skewed, some being rarer than others (Krawczyk, 2016: 221; Weiss, 2013). Such imbalance can take the form of a *class imbalance* – a category (or multiple) appearing rarely within the data, or *within-class imbalance* where there exist rare sub-classes or sub-concepts that can easily be missed when attempting to delineate a category (Weiss, 2013: 16-17).

[2] One statistical study on training set sizes for classifiers describes a range between 80 and 560 annotated samples for most models, depending greatly on use cases (Figueroa et al, 2012).

Imbalanced data poses a challenge for supervised learning for two main reasons.[3] First, as mentioned above, random sampling will naturally sample majority categories more (Miller, Linger, and Mebane, 2020; Krawczyk, 2016). Practitioners will need multiple rounds of sampling to locate enough minority cases to generalize from. In cases of extreme imbalance, such minority categories might not even show up at all. Second, imbalances can affect the algorithm's learning. Without sufficient data on minority classes, the algorithm might fail to learn to distinguish between them and end up simply guessing the majority class for each new instance (Krawczyk, 2016).

A variety of solutions have been proposed for working with imbalanced datasets. Broadly speaking, we can distinguish between those applied in the annotation stage and those applied post-annotation. Annotation stage strategies refer to solutions applied before the full training set has been compiled and labeled (either before or during the annotation process), and include methods for choosing samples for coding without resorting to inefficient rounds of random sampling. Post-annotation strategies refer to those applied at the model training stage, once a training set has already been collected and labeled. We will review the post-annotation strategies here, before focusing on the annotation stage.

Post-annotation methods can be placed in one of two classes – those at the data level and those at the algorithmic level. At the data level, researchers can tune the dataset itself in order to mitigate the imbalance by artificially creating a balanced class distribution (Kaur, Pannu, and Malhi, 2019; Tyagi and Mittal, 2020). This can be done in two ways. First, a researcher can use

[3] Imbalanced data is an issue for unsupervised methods as well (e.g., Weiss, 2013: 19; Yousefi et al, 2016), yet we focus in this paper on supervised learning applications. Interestingly, some practitioners advocate utilizing unsupervised learning as a preliminary step to identify distributions and imbalances within an unlabeled corpus (Krawczyk, 2016: 228).

undersampling – identifying and removing very similar instances within the majority class (e.g., Tyagi and Mittal, 2020: 210). Second, a researcher can *oversample* – either by replicating minority-class samples, or by generating synthetic samples of minority categories via interpolation (Chawla et al., 2002; Tyagi and Mittal, 2020; Weiss, 2013).

As opposed to data level solutions, algorithmic level fixes leave the dataset as is, instead tweaking the learning algorithm to improve the learning of minority classes (Krawczyk, 2016: 222; Weiss, 2013). This can be done by adopting weighting schemes that include higher penalties for mistakes in predicting minority class instances (e.g., Sun, Kamel, and Wang, 2006; Tyafi and Mittal, 2020: 210). This pushes the learning algorithm to focus on learning features that distinguish the minority classes. An additional algorithmic level strategy is to utilize ensemble learning or boosting – training multiple classifiers on the data to then be combined into a single, comprehensive classifier (Liu et al., 2022). Such a set-up can help identify class distributions (for weighting) or help identify more robust classification boundaries between the classes (e.g., Liu et al., 2022; Sun, Kamel, and Wang, 2006).

Active Learning and other Annotation Strategies

While post-annotation strategies attempt to ameliorate the detrimental effects of an imbalanced dataset after the fact, annotation stage strategies attempt to steer the annotation process towards a greater balance of categories by oversampling samples likely to belong to the minority classes. The two approaches are thus logically independent and can be utilized in tandem. Here we discuss the most widely used annotation method, *active learning*, as well as *guided search* – another noteworthy method utilizing human domain expertise within the sampling process, before concluding with our proposed annotation strategy.

In active learning, the researcher begins the annotation process with the traditional random sample of the target corpus[4]. This sample is then given labels by a human coder, and then used to train a preliminary classification model. The basic premise is that with this initial model, the algorithm can point to data instances that will most improve its learning after being labelled and added to the training set. This basic procedure can be repeated multiple times, with the re-trained model continuing to suggest instances for labeling while expanding the training set (Miller, Linger, and Mebane, 2020; Settles and Craven, 2008).

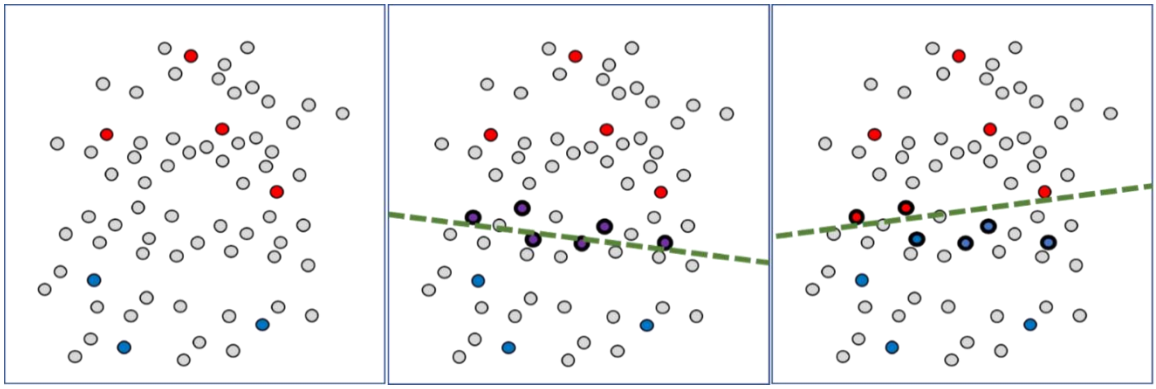
There are three major types of active learning – three approaches to deciding which data instances, when labeled, will most improve the learning process. The first is *uncertainty sampling.* In this method, the preliminary model is used to predict categories for unlabeled documents, noting the model's certainty of each prediction. After each such classification stage, the researcher can choose the most uncertain instances for human annotation, with the presumption that annotating these documents will best help the model learn to distinguish between categories. For simple classifiers, classification involves finding a single hyperplane (i.e., a plane that separates the full, multidimensional space into two halves) to distinguish between the classes. Thus, instead of sampling documents randomly, documents closest to the classifier's hyperplane are chosen for annotation. These new samples are then added to the growing list of labeled training samples, before a new intermediate classifier is trained and the process repeats. In this way, the computer can slowly retrain and develop a model through multiple iterations of human labeling of instances it is least certain about (see Figure 1 for

[4] The procedure can be implemented on a corpus represented computationally by any document embedding method, as long as the texts are represented in a vector form (much as in the way that an unsupervised clustering can run on any type of document representation scheme chosen by a researcher). Researchers can choose bag-of-words, pre-trained sentence embeddings, or even a Doc2Vec algorithm – depending on their research considerations.

illustration of uncertainty sampling). There are however several ways of measuring model uncertainty which may in practice differ.

Figure 1

Active Learning via Uncertainty Sampling



Note. In the first illustration (left), we can see a group of observations (for example, documents). A sample of those observations are then passed on to the researcher who annotates them based on their category (in this case, colored red and blue). In the next stage (center), a preliminary classification model is trained on those instances based on a single hyperplane, colored in green. A subset of those instances closest to the preliminary hyperplane (colored in purple) are then returned to the researcher for labeling. The researcher 'returns' the instances with their correct labels which is then used to train a new classification model (right). This process can be repeated multiple times to tweak the model as necessary.

The second type of active learning, often called *query by committee*, focuses on uncertainty as resulting from the disagreement of different types of models, rather than a single model's uncertainty (Settles, 2012: 21-35). The basic procedure is similar to that of uncertainty sampling – running multiple iterations of intermediate classification to identify uncertain data. However, here, at each step, multiple classifiers are trained on the existing training set, allowing

the algorithm to note instances of disagreement – documents or data points that the models are not consistent in their labeling (Miller, Linger and Mebane, 2020; Settles, 2012). Instances in which different models produce different results are sampled for annotation.

The final type of active learning utilizes *expected model change*. Here, at each step, a single model is trained. Then, the model estimates which data point, if labeled, would lead to the greatest change in the model's parameters or predictions (Miller, Linger, and Mebane, 2020: 6). While the algorithm may not know the true values for unlabeled instances, this approach utilizes *probability decision theory* to ascertain the *expected values* for each unlabeled instance (Settles, 2012: 37-38). Of these three approaches, the expected model change produces the most accurate classifiers with less labeled data, however, it is also the most computationally expensive (Settles, 2012: 38). Uncertainty sampling is the cheapest of the methods, and therefore, is the most commonly applied (Miller, Linger, and Mebane, 2020).[5]

A second annotation strategy is *guided search*. This method was introduced by Attenberg and Provost (2010) who demonstrated that active learning can be quite susceptible to initial bias. In an extremely imbalanced data set, skewness in the core training set can "lock in" an active learning process, missing key elements of categories, and push the entire learning process off course (Attenberg and Provost 2010; 2011). For example, in an uncertainty sampling procedure,

[5] In general, active learning is considered an expensive and cumbersome strategy, due to its requirement of multiple iterations of model training. Additionally, the final classification model must comply with the representations and models used at each of the intermediate classification stages (Lowell, Lipton, and Wallace, 2019). These requirements have constrained the utilization of active learning in conjunction with increasingly sophisticated and complex deep learning language models that are increasingly being utilized in computational communication applications. This is due to the inherent expenses and training times of such deep learning methods – which would be compounded over multiple iterations of the procedure. Such a limitation is one of several that we discuss.

the location of the initial hyperplane can vary greatly based on the first batch of samples and may continue to influence the position even over multiple iterations.

With this issue in mind, they propose performing annotation by utilizing the annotator's domain expertise to actively search for instances belonging to each category. In guided search, researchers create explicit strategies to target texts for sampling and labeling. This can be as simple as using a keyword search to search and filter documents. This ensures the inclusion of training samples for all relevant categories – especially those that are an extreme minority class.

These two annotation strategies offer sound options for researchers exploring rare categories in an unlabeled target corpus. However, there are several noteworthy limitations and biases we note before describing our proposed method. As mentioned earlier, active learning is vulnerable to an initial bias – locking into a learning trajectory due to the location of the initial classification boundary (Attenberg and Provost, 2010).

Furthermore, as the procedure continues, the method is susceptible to an additional *borderline bias* when choosing additional documents for human annotation. Active learning focuses on the most difficult instances lying at the margins between categories, to slowly adjust the boundaries. Thus, 'hard' cases, in which it may be difficult to distinguish between categories, are oversampled. However, the most uncertain instances may not necessarily be the most informative. Several studies have shown that active learning ends up focusing on outliers in the data, while missing out on instances that are actually representative of target concepts (Hu et al., 2010; Karamcheti et al., 2021). Thus, essential elements of target categories can be overlooked, and a focus only on hard cases may turn out to be unproductive (See Monarch, 2021: 8-9).

While such borderline bias can be problematic in any classification task, such issues are compounded when attempting to classify social science categories. Many theoretical concepts of interest may be complex, amorphous, and difficult to delineate (Kantner and Overbeck, 2020). While targeting of the margins is logical if we presume explicit differentiation between categories, this may not be the case for social science concepts, where true conceptual fuzziness is to be expected. Labeling such peripheral documents may be a nearly impossible task for human labelers. This could be an inefficient use of expert efforts, necessitating lengthy deliberations over difficult cases' labels, and producing more inconsistent annotations from different annotators, as well as lower inter-coder agreement rates (even in cases of multilabel classification, human coders may still not agree on what labels to assign). Thus, rather than efficiently producing a training set of representative texts that capture the core meaning of a concept, active learning may collect unclear instances, hampering the learning process and generalization.

Guided search is vulnerable to its own biases. The method is based on manually assembled keywords in creating the training set and is thus strongly influenced by idiosyncrasies of term choice. Studies have demonstrated that even domain experts can have difficulty in assembling accurate keyword searches for complex queries (King, Lam, and Roberts, 2017), especially with amorphous or complex categories like those found in the social sciences. Attempts to formulate definitions and guidelines at the outset of research can be vulnerable to researcher bias and incomplete conceptualizations (Collier, 1995). Additionally, while active learning contains an inherent borderline bias, guided search may tend to hone in on clear-cut, 'easy' instances, that belong to the category at hand in virtue of containing simple keywords.

Injecting human expertise efficiently into the annotation process

So how can we modify annotation strategies like active learning and guided search to help computational communication researchers attempting to handle the rarity and inherent complexity of our field's theoretical concepts? In this paper, we outline an annotation stage strategy for selecting texts for labeling which helps mitigate the issues of imbalanced data in supervised applications for social science research. Specifically, we attempt to utilize the strengths of both previously described methods, while addressing the issues of bias.

We look to the promise of interactions between humans and automated learning models to advance computer-assisted research, seeking to expand the foundations introduced in guided search. Such symbioses can mitigate human bias (King, Lam, and Roberts, 2017), ensure the learning algorithm's validity (Attenberg and Provost, 2011, p. 40; Grimmer and Stewart 2013), allow for the combination of quantitative and qualitative methods to ask theoretically significant questions (Baden et al., 2020; Jungherr and Theocharis, 2017, p.104; Ophir, Walter, and Merchant, 2020), and even improve the efficiency and accuracy of machine learning models (Monarch, 2021).

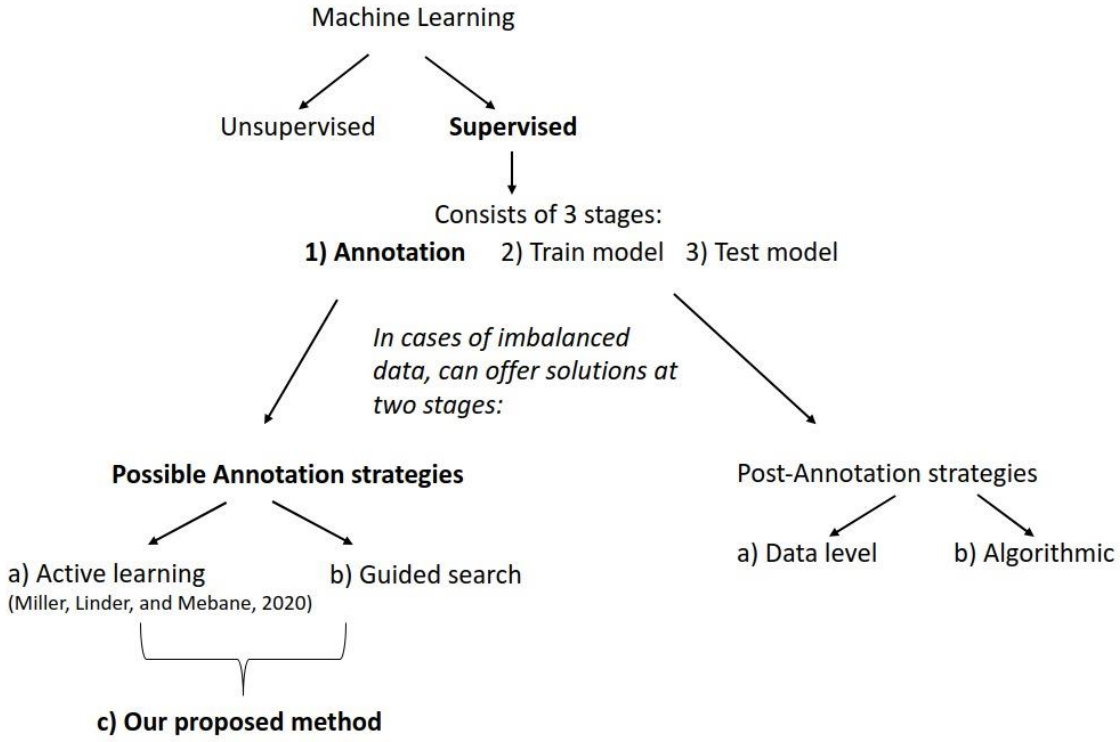
Thus, we seek to utilize researcher domain expertise, yet without being confined to narrow conceptualizations of our target categories due to the idiosyncrasies of our human coders. Our proposed method therefore seeks to combine domain experts' ability to explicitly target example texts with a targeted sampling method, capitalizing on their annotations in conjunction with features of the text detected in an unsupervised manner to target other relevant documents at lower cost, while avoiding and correcting researcher biases. Utilizing stratified sampling from a

latent document space, our method selects instances for annotation likely to belong to the target categories.

We specifically keep an eye on our use of human interaction throughout our procedure, prioritizing the utilization of such effort and expertise efficiently. We incorporate such human effort in two ways. First, in the initial targeting of documents to begin the annotation process. Here we are inspired by the guided search approach, utilizing domain expertise in choosing documents in line with the operationalization of a target concept. This interaction is relatively costly, but with high returns (as we will demonstrate using computerized simulations). Next, we continue to utilize researcher effort to label subsequent sampling rounds. However, our method attempts to limit the strain on human coders, offering relatively simpler labeling tasks than those used in active learning.

Figure 2

Overview of Theoretical Background



1

Note. This flow chart provides an overview of the approaches and methods described in the review above. At the top, we note the two primary machine learning approaches – unsupervised and supervised methods. In this paper, we focus on the supervised approach – which includes three stages: annotation, using the annotated training set to train a model, and then testing the model on a test set of samples labeled and external to the training set. We focus on the annotation stage, examining the issues in annotation on an imbalanced data set. While there are a variety of solutions to mitigating category imbalance in a dataset, in this manuscript we focus on the annotation stage. Specifically, we combine the advantages of two such strategies – active learning and guided search – into our proposed method.

Method

With these considerations in mind, we outline here a method for selecting documents from an unlabeled corpus for annotation that utilizes domain expertise at key junctures for a more comprehensive sampling strategy. Our contribution is in the combination of an expert targeting, together with the latent semantic vector space of an unlabeled target corpus to accumulate additional instances for annotation while mitigating initial researcher bias and error. Specifically, we embed the documents of our unlabeled target corpus in a shared multidimensional *semantic space*. We then use preliminary researcher input to locate starting points for each concept within the shared semantic space. Then, we implement stratified sampling of documents from their surroundings. This allows us to target texts most likely to be related to and representative of each concept, while still exploring the uncertain peripheries to attempt to ascertain the best classification boundaries and correcting possible biases and limitations in domain experts' initial targeting. In this way we allow researchers to utilize the advantages of active learning while avoiding the pitfalls of borderline bias. Additionally, we make use of human input in an efficient manner – focusing such efforts not only on difficult, inconclusive samples, but on a continuum of instances to both delineate and represent the theoretical concepts of interest. We outline the steps below, before demonstrating the full procedure's viability in multiple simulations. The full list of steps can be seen in Table 1 at the end of this section.

Creating the base nucleus for each target concept

We begin our procedure with the first stage of human domain expertise – collecting the base nucleus of documents for each concept. Such nuclei are the group of core texts representing each concept and will be the basis of the forthcoming unsupervised expansion detailed below. This nucleus consists of a small set of documents chosen by a domain expert to best represent each concept, based on a preliminary definition of said concept. This nucleus should include a

heterogeneous list of texts to best capture the multiple facets and complexities of the theoretical concepts. For example, in a study attempting to classify political identities such as *Liberal* and *Conservative*, the researcher should attempt to collect documents covering such concepts from as many angles as possible – including stylistic elements and various thematic domains[6]. This can be done via domain knowledge, metadata or even by simple keyword searches. For example, a researcher exploring political ideologies might utilize her expertise to search out relevant sections of party manifestos or editorials from partisan news outlets regarding specific issues. Such documents need not actually be found in the target corpus, as long as they are similar types of texts (such as social media posts or news articles).

Creating document-level embeddings and centroids

After having compiled the original group of core documents for each concept, we generate document-level embeddings to represent statistically each one of the labelled nucleus texts, as well as the full pool of unlabeled documents in our corpus of interest. The basic concept of document embeddings is the computational representation of texts into vectors – a list of numbers – that can then be plotted within shared, multidimensional space based on their semantic content.

This shared semantic space allows us to measure distances between vectors corresponding to the similarities between each data point. The concept of a shared semantic space is utilized frequently when working with word embeddings. The geometric distances

[6] As noted previously, supervised classifiers can be tailored to measure diverse types of categories, allowing researchers to choose what operationalization is most relevant for their study. For example, a researcher might hone in on documents written by Conservatives, discussing Conservative issues or expressing specific sentiment to Conservative stances, depending on the research question. Additionally, the flexibility in design means that researchers can consider documents with multiple labels or even nulls when compiling such nuclei.

between word vectors have been revealed to correspond to semantic relations (e.g., Bolukbasi et al., 2016; Mikolov et al., 2013). For example, words with similar meanings are found within closer vicinity to each other than unrelated words (See Figure 3).

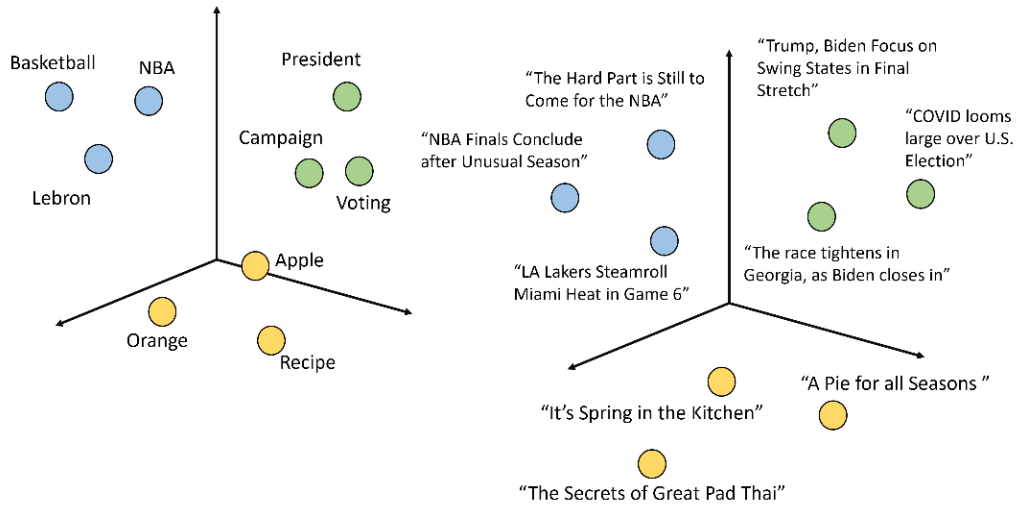
We can utilize these geometric distances to reveal relationships between document-level embeddings as well. The closer the distance between a pair of documents in this space, the more they are similar semantically. For example, in Figure 2 we can see an illustration of the mapping of document-level vectors for each category. Such distance-relationships will form the basis of our unsupervised exploration of the target corpus, allowing us to quantify the similarity between documents and search for additional samples accordingly.

While several approaches to embedding documents exist (e.g., Corrêa Júnior, Marinho, and Santos, 2017; Zhao, Lan, and Tian, 2015), here we have opted for embeddings from a pre-trained SentenceTransformers model to encode the sentences in our documents. Such models are pre-trained on semantic textual similarity (STS) tasks and produce an embedding vector representing the contextual meaning of the input sentence, so that semantically similar sentences receive similar embeddings. While different embedding schemes may emphasize different textual characteristics and may better suit specific applications, sentence embeddings based on semantic similarity provide a widely applicable embedding space, as we shall shortly discuss. After encoding the sentences into vectors, we calculate the average of the sentence embeddings (mean of the group of vectors) to find a single vector to represent the full document (Reimers and Gurevych, 2019). These unsupervised document embeddings are then used to select additional texts for annotation.[7]

[7] We provide a demonstration of the use of Sentence Transformers in Appendix B.

Figure 3

From Word to Document Embeddings

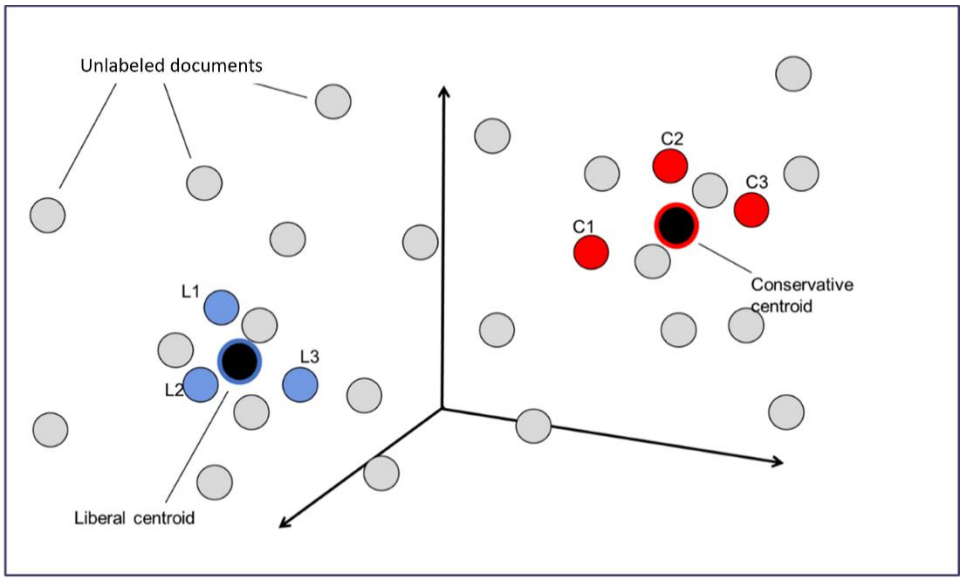


Note. The graph on the left illustrates the geometrical distances between word embeddings, with more similar words found closer to each other (colors added to highlight thematic clusters). The graph on the right illustrates the same phenomenon at the document level. In this example, a corpus of news articles has been converted to document-level embeddings. Each article is represented by a colored circle with the corresponding title displayed. Here too, colors have been added to highlight the closer distances based on the similarity between documents.

With this understanding, we can return to our method. At this stage, we have our collections of representative documents for each concept of interest, as well as the wider target corpus. We then calculate document-level vectors for all texts, utilizing the same procedure described above. For each base collection, we then calculate the mean vector of the documents, finding a centroid that captures the average semantic meaning representing the concept. Figure 4 illustrates the finding of centroids within a shared semantic space.

Figure 4

Illustration of a Shared Semantic Space



Note. Each circle represents a document's embedding within the space. Similar documents are found within closer proximity of one another. In this example, the blue and red colored circles correspond to the core documents chosen by the researchers to represent the concepts of 'Liberal' and 'Conservative' (respectively). The blackened circle represents the centroid calculated for each concept. These centroids form the basis around which the unsupervised expansion will occur. Finally, the grey circles are the additional documents of the corpus that are still unlabeled, having not yet been determined if they express a target.

Expanding the training set through stratified sampling

Having calculated our centroids, we now will utilize them to begin a comprehensive sampling of the shared semantic space. Bearing in mind the need for both borderline documents such as those

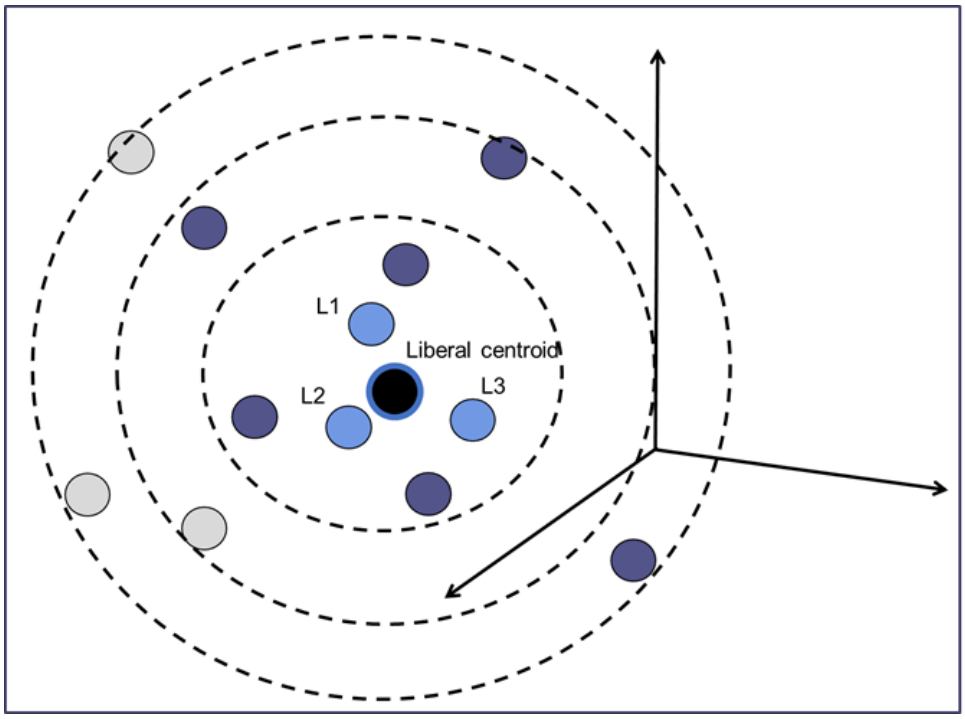
targeted in active learning, and core documents that represent concept depth, we now perform a stratified sampling for each target concept randomly choosing documents from each hierarchal 'strata' of receding cosine distance from the centroid. Thus, we are ensuring a wider, more heterogeneous collection of samples for a training set – attempting to expand our representation of and encompass fully the target concepts.

To clarify, documents located at a 0.8 cosine distance from the centroids should be closer conceptually to the corresponding concept (a cosine distance of 1 signifies no distance - i.e., total similarity), while documents found at a 0.5 cosine distance are expected to be less relevant. These samples can then be passed on to human coders who, oblivious to the distances of the document from the centroid, mark which ones are related to the said concept (see Figure 5). After each such sampling round, the documents labeled as relevant by the coders can be added to the existing lists of samples per each concept, and a new centroid can be calculated from this growing collection of texts.

While such multiple rounds of sampling across the spectrum of distances and certainty for each target concept should provide a robust sampling strategy, we note the importance of including 'sanity-checks' between each iteration as a validation that the method is working as expected. Specifically, following coders' labeling of the stratified samples, we would plot their coding results in graphs. In these graphs, we would sort the samples deemed to be relevant to each concept by the human coders based on their distance from their corresponding concept's centroid. Such graphs would offer opportunities to take a glance at the procedure's progress, validating its utilization.

Figure 5

Illustration of Stratified Sampling for Training Set Expansion

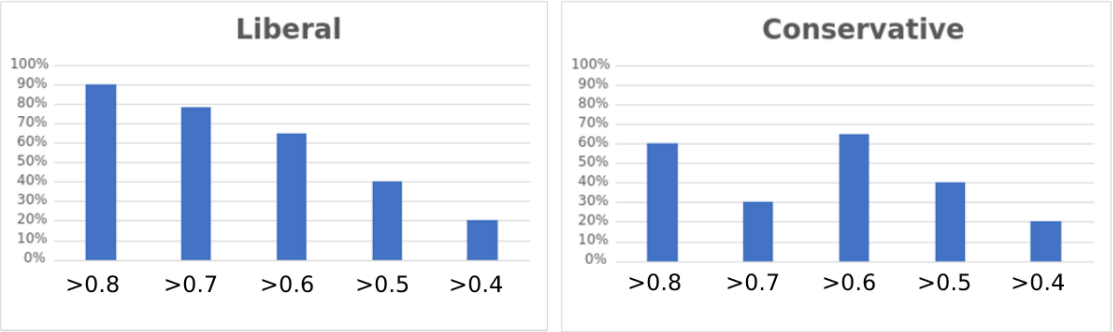


Note. Here we see the base documents (light blue circles) surrounding the centroid (black circle). We extract a random sample of documents found at receding distances from the centroid, and human coders label whether the documents are related to the concept (dark blue circles) or unrelated (grey circles).

For example, the graphs in Figure 6 show the percent of randomly sampled documents deemed to be relevant to a said topic from each hierarchal level of cosine distance from the centroid in a hypothetical study examining the concepts of “Liberal” and “Conservative". On the x-axis, the left most column shows the percent of documents found relevant between a 0.8 and 0.9 distance from the centroid, while the right most column shows those found relevant between a 0.4 to 0.5 distance. In the graph on the left, we see the results of what we assume a coherent concept looks like, in this case, 'Liberal' – we can see the overall decline of relevant documents as we move away from the centroid.

Figure 6

Stratified Sampling Results for Two Concepts



However, if we do not see a monotonically decreasing function as we move farther from the centroid, we can infer that the method is not functioning as we had hoped. This could be due to several possible issues. A first possibility might be problems with the choice of the base documents used to create the nucleus for each target concept. Perhaps the samples chosen to best represent each category do not comprise a coherent operationalization of the theoretical concept. Our field seem particularly prone to such issues. As mentioned earlier, many social science concepts can be 'soft' - complex, multidimensional, and difficult to delineate. For example, the concepts displayed here – 'Liberal' and 'Conservative' - span multiple thematic categories and stylistic dimensions. Any attempt to empirically explore such concepts can be vulnerable to researcher bias in the initial operationalization of such identities. In defining such concepts, a researcher might miss out on crucial elements inherent in them. Such elements are those that the researcher may not have conjured up when defining the concept at the start of the study, yet, upon seeing a document expressing said facet, realize that it is integral to the concept. In such cases, the graphs of our stratified sampling offer an opportunity to uncover unidentified theoretical dimensions and expand our concept.

For example, in the graph on the right in Figure 6, we see the stratified sampling for the concept of 'Conservative'. Here we see an indication for an unidentified theoretical dimension – instead of a monotonically decreasing function as we move farther from the centroid, we see a slight increase in relevant documents at a distance (*> 0.6 strata*). We can examine such documents to ascertain if they truly encompass a relevant element or dimension. If so, we can then add those newly-labelled documents to our growing collection of relevant samples, recalculate a new centroid, and run another round of stratified sampling. Thus, we can iteratively continue adding new, labelled instances that cover multiple dimensions of our target concepts.

If, however, we see few relevant documents across the hierarchal strata, we can infer that the sampling process is lacking – either due to the choice of embedding scheme or even the initial operationalization of the concept as reflected by the centroid location in the shared semantic space. We can then attempt to tweak the procedure – either by utilizing different embedding models for our documents or by adding additional core documents to try a different initial centroid location. This feedback can be used in various ways to improve the annotation even at the start of the procedure, or throughout any of the ensuing steps.

A final possibility is that the distances chosen for the strata may be unsuited for the target concept. For example, some clusters may be smaller and condensed, while others may span larger swathes of the semantic space. In such cases, we may see graphs with extreme distributions or skewness that could point us to use different strata distances.

Crucially, we note that most issues described above will be explicit by the first stage, allowing the researchers to solve them quickly. Moreover, such improvements in the representation of texts via embeddings or the choice of core documents will not simply ‘fix’ the

annotation process but will also contribute to better performance of the final classification model. Thus, such effort could have a significant influence on the entire research project.

Training the Classifier

As is usually the case in supervised learning, there are no explicit rules of thumb as to when to stop annotating documents and to move on to training the final classification. The customary method of determining this is by examining model accuracy (either on a held-out validation or test set, or via cross validation) and checking whether performance is satisfactory. While adding additional annotated documents will generally lead to continuous improvement in predictive performance, its returns are diminished over successive iterations.

However, our proposed method provides a useful lower bound via the movement of centroids. After each round of stratified sampling, a new centroid is calculated to include the new, labeled samples. As long as these centroids continue to move within the semantic space, this indicates that the concept is still 'raw' - collected documents may not yet be representative of the concept at hand. When centroids become stable, or movement is negligible, it may be time to examine model performance. At this stage, the researcher can choose to use any type of classification – be it SVM, Naïve-Bayes or even fine-tuning a pre-trained deep learning language model via transfer learning. Additionally, we note that in our procedures researchers qualitatively examine the training set at each iteration. Should they feel confident in their collection at any round, they can train an intermediate model, check the classification on the test set, and determine whether the model is reliable enough.

Table 1

Procedural Steps for the Proposed Semi-supervised Method

Step	**Description**	**Human Input**
Creating base nucleus for each concept	Collecting group of documents by domain experts around which to expand the delineation of target concepts in the full corpus.	Domain expertise in operationalizing theoretical concept, hand-picking documents to best represent target category.
Creating document-level embeddings and find centroids	Allows the mapping of the shared semantic space of the corpus.	
Stratified Sampling	Utilize stratified sampling to expand a concept along a continuum, creating depth in covering documents with varying levels of association and uncertainty for each target category.	Human coders label samples without knowing distance from centroid. Researchers examine validation graphs to inspect procedure's efficacy.

Train the supervised classifier	Use the iteratively collected groups of documents as the labelled training set for a classification model.	

Simulation

In order to evaluate the efficacy of our proposed method, we designed a number of simulations allowing us to compare it to the two most prominent annotation strategies, serving as baselines: random sampling and active learning. The use of simulations allows researchers to run multiple iterations of a proposed method under diverse parameters, helping to ascertain the efficacy of such procedures under diverse conditions (For example of use of simulation, see Miller, Linger, and Mebane [2020]). We decided to run such tests on two existing, fully annotated datasets of different text types: the 20 Newsgroups collection[8] - a popular dataset of internet newsgroup posts (similar to Reddit) that is a classic resource for many machine learning applications and studies, and the New York Times Front Page Dataset (Boydstun, 2013)[9] – a compilation of newspaper front page headlines that have been classified within the framework of the Comparative Agendas Project (CAP) and utilized for political science research (Dowding, Hindmoor, and Martin, 2016). We extended the New York Times dataset by extracting the full news articles from LexisNexis. We have attached the descriptive statistics and list of categories for each dataset in Appendix A.

We utilized two separate simulation designs: the first to quantify the efficiency of our sampling strategy on different data distributions, and the second to demonstrate the practical feasibility of the tasks given to the human coders. In the first design, we utilize the ability to run thousands of iterations of the sampling procedures to compare the accuracy of random sampling, active learning, and our method over varying category distributions in a systematic manner. Additionally, to bolster this evaluation, we ran such simulations 300 times over each of a number

[8] http://qwone.com/jason/20Newsgroups/
[9] Full data taken from http://www.amber-boydstun.com/supplementary-information-for-making-the-news.html

of category distribution schemes, while also randomly choosing the target categories so as to ensure that our final average accuracy levels for each strategy would be due to the approach chosen, and not to idiosyncrasies of category choice or other inherent random elements. For example, in our simulation on the New York Times Front Page Dataset, for each round we might choose a different macro category (from "Health", "Defense", etc.) as the target category to classify among the aggregated others. (We offer a comprehensive description of the simulation procedures below.)

This form of simulation has the advantage of allowing us to examine a host of very different settings – datasets and distributions of categories – and increasing generalization capacities by thousands of repetitions. However, the cost of this is that the human element is simulated away. Systematic examinations would not be possible with human coders, as they necessitate tens of thousands of iterations of our procedure and on different category schemes and datasets. Instead, the texts chosen for annotation in each sampling round are simply assigned the true label given in the dataset.

In order to ascertain the practical feasibility of the tasks given to our human coders, we chose one classification design to run fully, with human annotators, in a second simulation stage. We chose to focus on a single classification task – choosing three minority classes from the 20 Newsgroups collection: Baseball (812 posts), Hockey (809 posts), and Middle East Politics (792 posts), while aggregating the rest into a single 'Other' category (12,696 posts). The two sports categories were chosen to evaluate the classification on two similar classes, while Middle East Politics was added to provide a vastly different type of category. We then implemented the stages of human input, instructing our coders to first compile the base nucleus of texts for each category using keyword searches to find relevant posts. We then ran two rounds of our stratified

sampling procedure, handing over the chosen posts to our coder for labeling. After each such round, we trained an intermediate SVM classifier, noting the accuracy levels on the test set. (The full description and coder instructions for this simulation can be found in Appendix C.) Finally, we ran the fully automated simulation from the first stage, focusing only on the three categories utilized in the human-input simulation. This way, we ensured that our method outperforms the alternative annotation strategies in this particular classification task.

The basic format of the first simulation design is as follows:

(1) Texts were embedded in a latent semantic space by utilizing sentence embeddings from pretrained language models optimized for semantic similarity (Reimers and Gurevych, 2019)[10]. Such models assign to a sentence or short text a dense vector that represents its linguistic and semantic features, such that similar texts would be assigned similar vectors. Sentence-level vectors are averaged in order to produce document-level embeddings. These vectors are utilized by the proposed sampling strategy and form the input data for models trained via all sampling strategies in this simulation. In our case, we utilized the Sentence-BERT model (Reimers and Gurevych, 2019).

(2) Create an imbalance in the data, choosing one category to under sample, thus increasing their rarity in the larger dataset. In simulations using the New York Times articles, we would choose a single domain from the multiple macro-domains of the CAP codebook, label the rest of the data as being 'Other', and then attempt to classify the narrower sub-

[10] It is important to note that there are many embedding schemes to choose from, and researchers can even train their own custom embeddings on the target corpus. Ultimately, in our use-case here – the simulations – the choice of embedding scheme is inconsequential as any inherent biases or deficiencies of the document representations are uniform across the three methods compared: random sampling, active learning and our method.

topics. A test set representative of the imbalanced distribution is randomly selected and put aside. We note that researchers in their own applications can choose to manipulate the distributions of categories within their own test sets, depending on the theoretical considerations of their use case as well as learning algorithm choice.

(3) Then, compile a core sample training set, completely random for random sampling and active learning, and distributed among all target categories for our method – corresponding to the role of human experts in that procedure. (This corresponds to Step 1 in Table 1.)

(4) Train a SVM classifier on each of these training sets and record the accuracies for each on the test set.

(5) Continue expanding the training sets for each category through multiple iterations of sampling based on each approach: random sampling, choosing the most 'uncertain' instances for active learning, and a stratified sampling for our method (in our simulation, we randomly chose one target category per stratified sampling round to keep the training sets equal between the three approaches). Whereas in a real-world utilization of such methods, human coders would manually label such samples, in our simulation we simply used the true categories of each sample as the labels. (This corresponds to Step 3 in Table 1.)

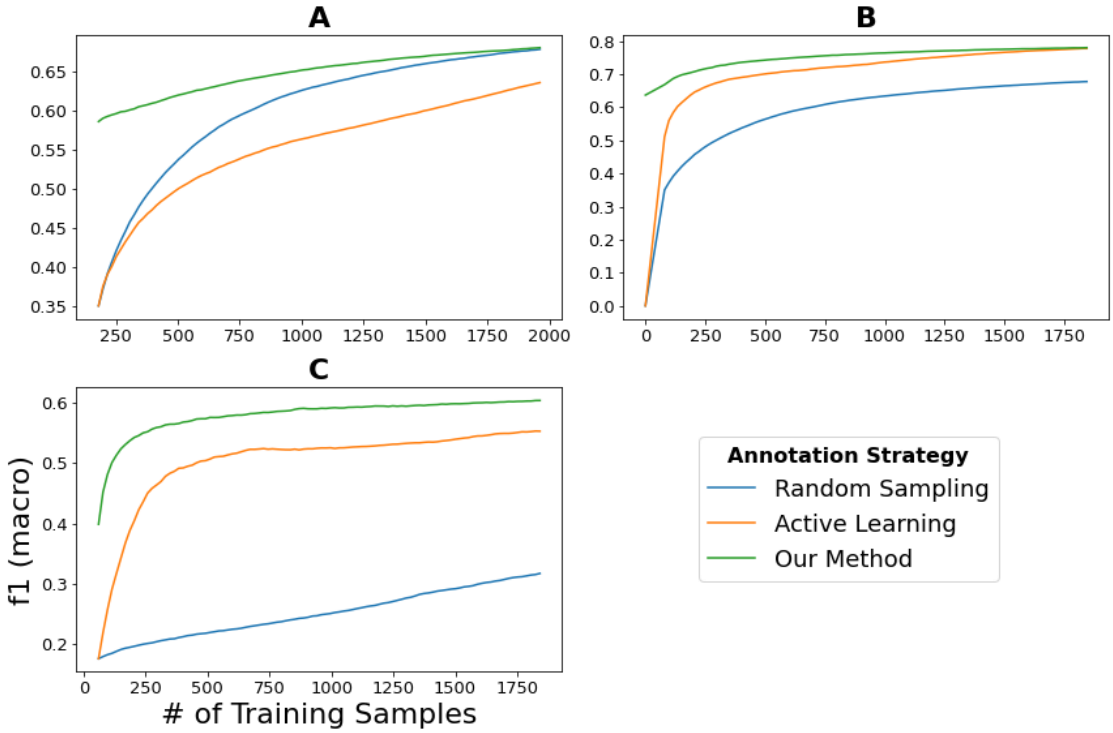
(6) After each such sampling iteration, we would train a SVM classifier on the expanded training sets, recording the accuracy of each model. (This corresponds to Step 4 in Table 1.)

Results

We ran these simulations over several category distribution schemes, to ascertain the comparative efficiency and accuracy of our proposed annotation method. Specifically, we found that our method consistently compiled initial training sets allowing for significantly more accurate predictive models, and that in many additional schemes, such leads were maintained over multiple sampling rounds (See results in Figure 7).

Figure 7

Simulations Comparing Proposed Method to Random Sampling and Active Learning



Note. On each x-axis, we see the number of samples in the training sets, slowly accumulating more annotated documents as we continue the sampling iterations. On each y-axis, we see the F1 accuracy scores of the intermediate classification models. The three strategies are marked by color: blue for random sampling, orange for active learning, and green for our method. In Graph A, we see a simulation using the 20 Newsgroups data, in this case utilizing the original nine categories in full in a relatively balanced dataset. In Graph B, we randomly chose one category to

under sample, and three categories to keep at their original size – thus comparing these approaches in cases of category imbalances. In Graph C, we see a simulation on the New York Times Front Page dataset. For each run, we randomly chose a single macro category from the CAP coding scheme (e.g., "Defense", "Health", "International Affairs and Foreign Aid", etc.), aggregating all the other categories to a single category ('Other'), and then classifying the sub-categories within the larger, unrelated dataset. This best emulates routine use-cases in political science research – trying to identify theoretically-relevant, politically-oriented concepts, within a larger database mostly unrelated to the research question.[11]

Additionally, in order to supplement these automated simulations, we ran a shortened version of our procedure to demonstrate the feasibility of the human element in our method. After compiling the base nucleus of posts and running two rounds of annotation based on our stratified sampling, we achieved high levels of accuracy (Full results in Table 2). More importantly, we achieved such results with our human coder having read only 284 texts (consisting of informal, mostly short, internet posts) in approximately 4 hours. Such effort seems entirely reasonable and cost-effective for researchers looking to employ our method in developing a high-quality, theory-driven classifier. (The full list of sampled posts and their labels is found in Appendix D.) Additionally, we ran a full computerized simulation comparing our method to random sampling and active learning on this category scheme. In Figure 8, we can see the advantage of the stratified sampling over the other methods over multiple iterations.

[11] We direct the readers to an interesting result in Graph A, where we ran our simulations on the original category scheme with relatively balanced data. We can see that in such a case, the random sampling outperforms the active learning approach, affirming that the focus on uncertain cases in such datasets is less efficient and reliable.

Finally, in order to ascertain the ability to combine our annotation method with the latest, state-of-the-art deep learning language models, we trained a RoBERTa transformer model on our labeled samples and ran on our test set. As we can see in Table 2, the RoBERTa classifier achieved impressive F1 accuracy levels utilizing our method. This demonstrates powerful applications of our method for researchers – an efficient annotation procedure, that, when combined with the latest language models, provides potent classification capabilities.

Table 2

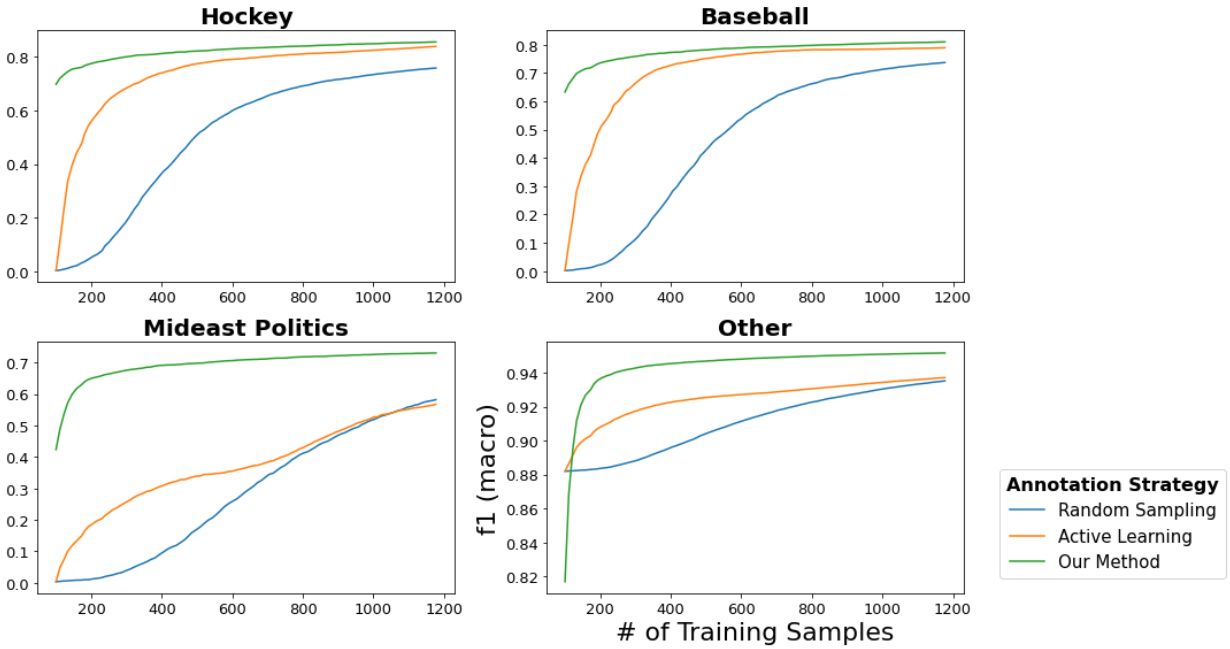
Accuracy Results after Rounds of Human Annotation

Category	First Stratified Sampling Round (F1 score)	Second Stratified Sampling Round (F1 score)	Second Stratified Sampling Round - RoBERTa model (F1 score)
Baseball	0.698	0.694	0.84
Hockey	0.755	0.770	0.81
Middle East Politics	0.606	0.677	0.73
Other	0.962	0.968	0.98

Note. We see here the results of a classifier trained on the training set after both the first and second rounds of stratified sampling in the second simulation. We see strong performance (F1 scores) after the first round for each of the classes, and further improvement after adding the second rounds. Additionally, when we trained a RoBERTa transformer model on the annotated samples, we reached significantly higher scores.

Figure 8

Simulations Comparing Annotation Strategies for three 20 Newsgroups Categories



Note. We see here the accuracy results for the three annotation strategies for each of the three categories chosen for the second simulation, together with the aggregated 'Other' category. We see here that for each of the classes, our method outperformed the alternative strategies.

Discussion

These simulations reveal several points worth dwelling upon in the procedure outlined in this paper. First, we note the importance of domain expertise at the start of any annotation process. We can observe the stark gap in the starting points between our method and the two others attempted here. Such differences are quite understandable, as the researchers can utilize their

domain expertise to craft balanced training sets covering all target concepts. This initial lead can be quite crucial. As described in the introduction, multiple iterations of human annotation can be quite costly. By reaching relatively high levels of classifier accuracy early on, we are able to significantly reduce the number of iterations needed to prepare a sound classifier.

However, while significant, the gaps between our method and the others are not attributed to the initial sampling alone. We can see from the graphs that our method maintains an edge over the others over multiple iterations of training set expansion. This validates the importance of stratified sampling. By not only looking towards the borderline, uncertain instances, instead exploring the full semantic space surrounding a target category we can improve the ability of a classification model to identify such categories within a larger corpus.

In addition to these improvements in classifier accuracy, our method provides a more efficient procedure than the others described here. While the initial stage is expertise-expensive due to the hand-picking of documents for the core collection, such effort is of great worth – proving itself in the large gap in accuracy in the first round of classification. Additionally, in the following rounds, the labeling of documents in our method is relatively simpler than for other methods, such as active learning. Labeling the 'borderline' instances – documents that are either categorically ambiguous or may even contain multiple themes within them – can be exceedingly difficult for human coders. By not only focusing on such difficult cases, instead offering a spectrum of samples to label, we are in effect reducing the deliberation and effort of our human coders. It is important to note that such nuances are not covered in our computerized simulations, where we utilize the true labels for each document without researcher involvement. Thus, efficiency gaps may be even starker in the utilization of the method 'in the wild'.

However, we stress that even if our method proves more efficient in particular research cases, we do not profess universal superiority or attempt to dissuade researchers from utilizing other strategies (such as active learning) in their own work. Such methods might even be used in conjunction with our procedure in ways that could maximize efficiency and accuracy for their own tasks. Identifying and understanding such combinations could be an interesting and promising future research endeavor.

Conclusion

With the supervised approach becoming more prominent in machine learning, the method proposed here builds upon existing techniques while utilizing expert knowledge in a more efficient way to offer significant returns in accuracy while compiling an annotated training set. We emphasize the focus of this paper – a method improving the annotation stage for supervised machine learning endeavors on an imbalanced corpus. Imbalanced data and rare categories are a significant challenge, and several techniques have been offered to mitigate such issues – both in the sampling stage by artificially balancing the data, or in the learning stage through weighting and algorithm choice. Our procedure does not constrain the researcher in any of these stages and is compatible with any technique the researcher should choose for the application of supervised learning in their study. Ultimately, our method here allows high level of researcher participation in the learning process, both utilizing domain expertise to efficiently tackle the issues of imbalance, as well as allow the researcher to track and validate the process.

While the thoroughness of the procedure may seem to be daunting and costly in time, in our experience, the brunt of the time was spent preparing the document-level vectors for the entire corpus. Yet this only needed to be done once, at the start of the study. As such, additional

concepts can be explored and added with little extra effort, based almost entirely on the pace of the human readers' validation, as was demonstrated in the second stage of our simulation. Additionally, the human effort is utilized quite efficiently – with most of it being used to create a preliminary training set (returning significant increases in accuracy), while being given simpler coding tasks at later iterations.

In addition to the efficiency of our proposed method, we offer here a qualitative edge which was not fully reflected in our simulation – the ability to handle complex, theoretically driven concepts. As described above, existing machine learning methods are limited in their ability to cover the full range of facets and criteria included in such theoretical concepts. Thus, we offer here a strategy for dealing with important concepts found throughout the social sciences. The development of new tools tailored to the requirements of social sciences can open new opportunities for empirical research of theoretical concepts.

Replication Materials can be found online at https://dataverse.harvard.edu/privateurl.xhtml?token=ec8787c1-1d3a-4f0c-b999-0cc21adaf4f4

References

Attenberg, J., and F. Provost. 2010. “Why label when you can search? Alternatives to active learning for applying human resources to build classification models under extreme class imbalance.” In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge discovery and data mining, 423–432. KDD.

Attenberg and Provost. 2011. “Inactive Learning? Difficulties Employing Active Learning in Practice.” SIGKDD Explor. Newsl. 12(2): 36–41. https://doi.org/10.1145/1964897.1964906.

Baden, C., N. Kligler-Vilenchik, and M. Yarchi. 2020. “Hybrid Content Analysis: Toward a Strategy for the Theory-driven, Computer-assisted Classification of Large Text Corpora.” Communication Methods and Measures. 14 (3): 165–183. https://doi.org/10.1080/19312458.2020.1803247.

Barberá, P., A. Casas, J. Nagler, P. J. Egan, R. Bonneau, J. T. Jost, and J. A. Tucker. 2019. “Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data.” American Political Science Review 113 (4): 883–901. https://doi.org/10.1017/S0003055419000352.

Bauer, M. 2000. “Classical content analysis: a review.” In M. W. Bauer, and G. Gaskell (Eds.), Qualitative researching with text, image and sound, 132-151. SAGE Publications Ltd, https://dx.doi.org/10.4135/9781849209731.n8

Bhattacherjee, A., 2012. "Social Science Research: Principles, Methods, and Practices." Textbooks Collection. 3. https://digitalcommons.usf.edu/oa_textbooks/3

Bolukbasi, T., K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. 2016. “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.” https://arxiv.org/abs/1607.06520.

Boydstun, A. E. 2013. Making the news: Politics, the media, and agenda setting. Chicago, IL: University of Chicago Press.

Chatsiou, K., and S. J. Mikhaylov. 2020. “Deep Learning for Political Science.” In The SAGE Handbook of Research Methods in Political Science and International Relations, 1053–1078. SAGE Publications Ltd. https://doi.org/10.4135/9781526486387.n58.

Chawla, N. V., K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. 2002. “SMOTE: Synthetic Minority Over-Sampling Technique.” Journal of Artificial Intelligence Research, 16: 321–57. https://doi.org/10.1613/jair.953.

Collier, D. 1995. “Translating Quantitative Methods for Qualitative Researchers: The Case of Selection Bias.” American Political Science Review 89 (2): 461–466. https://doi.org/10.2307/2082442.

Corrêa Júnior, E. A., V. Q. Marinho, and L. B. dos Santos. 2017. “NILC-USP at SemEval-2017 Task 4: A Multi-view Ensemble for Twitter Sentiment Analysis.” In Proceedings of the 11th International Workshop on Semantic Evaluation - SemEval-2017, 611–615. Vancouver, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-2100.

Dai, Y., and A. Kustov. 2022. “When Do Politicians Use Populist Rhetoric? Populism as a Campaign Gamble.” Political Communication 0 (0): 1–22. https://doi.org/10.1080/10584609.2022.2025505.

Dekel, O. and O. Shamir. 2010. “Multiclass-Multilabel Classification with More Classes than Examples.” In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 9: 137-144. https://proceedings.mlr.press/v9/dekel10a.html

Dobbrick, T., J. Jakob, C. - H. Chan, and H. Wessler. 2021. “Enhancing Theory-Informed Dictionary Approaches with “Glass-box” Machine Learning: The Case of Integrative Complexity in Social Media Comments.” Communication Methods and Measures 0 (0): 1–18. https://doi.org/10.1080/19312458.2021.1999913.

Dowding, K., A. Hindmoor, and A. Martin. 2016. “The Comparative Policy Agendas Project: theory, measurement and findings.” Journal of Public Policy 36 (1): 3–25. https://doi.org/10.1017/S0143814X15000124.

Figueroa, R.L., Q. Zeng-Treitler, S. Kandula, and L. H. Ngo. 2012. “Predicting sample size required for classification performance.” BMC Medical Informatics and Decision Making, 12 (8). https://doi.org/10.1186/1472-6947-12-8

Fogel-Dror, Y., S. Shenhav, and T. Sheafer. 2018. “Theory-driven Text Classification with Minimal Human Effort.” The 2018 Annual Meeting of the International Communication Association, Prague, 2018

Gilardi, F., T. Gessler, M. Kubli, and S. Müller. 2022. “Social Media and Political Agenda Setting.” Political Communication 39 (1): 39–60. https://doi.org/10.1080/10584609.2021.1910390.

Grimmer, J., and B. M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” Political Analysis 21 (3): 267–297. https://doi.org/10.1093/pan/mps028.

Guess, A., K. Munger, J. Nagler, and J. Tucker. 2019. “How Accurate Are Survey Responses on Social Media and Politics?” Political Communication 36 (2): 241–258. https://doi.org/10.1080/10584609.2018.1504840.

Jungherr, A., and Y. Theocharis. 2017. “The empiricist’s challenge: Asking meaningful questions in political science in the age of big data.” Journal of Information Technology & Politics 14 (2): 97–109. https://doi.org/10.1080/19331681.2017.1312187.

Kantner, C., and M. Overbeck. 2020. “Exploring Soft Concepts with Hard Corpus-Analytic Methods.” In Reflektierte algorithmische Textanalyse: Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt, edited by N. Reiter, A. Pichler, and J. Kuhn, 169–190. De Gruyter. https://doi.org/doi:10.1515/9783110693973-008

Karamcheti, S., R. Krishna, L. Fei-Fei, and C. D. Manning. 2021. “Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering.” CoRR abs/2107.02331. arXiv: 2107.02331. https://arxiv.org/abs/2107.02331.

Kaur, H., H. S. Pannu, and A. K. Malhi. 2019. “A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions.” ACM Computing Surveys 52 (4). https://doi.org/10.1145/3343440

King, G., P. Lam, and M. Roberts. 2017. “Computer-Assisted Keyword and Document Set Discovery from Unstructured Text.” American Journal of Political Science 61 (4): 971–988. https://doi.org/10.1111/ajps. 12291.

Krawczyk, B. 2016. “Learning from imbalanced data: open challenges and future directions”. Progress in Artificial Intelligence 5: 221–232. https://doi.org/10.1007/s13748-016-0094-0

Krippendorf, K. 2018. “Content Analysis: An Introduction to its Methodology.” Sage Publications

Lau, J.H., D. Newman, and T. Baldwin. 2014. “Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality.” Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 530-539. https://aclanthology.org/E14-1056

Liu, L., Wu, X., Li, S. et al. 2022. “Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection.” BMC Medical Informatics and Decision Making, 22(82). https://doi.org/10.1186/s12911-022-01821-w

Loftis,M.W., and P. B.Mortensen. 2020. “Collaborating with the Machines: A Hybrid Method for Classifying Policy Documents.” Policy Studies Journal 48 (1): 184–206. https://doi.org/https://doi.org/10.1111/psj.12245.

Lowell, D., Z. C. Lipton, and B. C. Wallace. 2019. “Practical Obstacles to Deploying Active Learning”. arXiv: 1807.04801 [cs.LG].

Merkley, E., and D. A. Stecula. 2021. “Party Cues in the News: Democratic Elites, Republican Backlash, and the Dynamics of Climate Skepticism.” British Journal of Political Science 51 (4): 1439–1456. ttps://doi.org/doi: 10.1017/S0007123420000113.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv: 1301.3781 [cs.CL].

Miller, B., F. Linder, and W. R. Mebane. 2020. “Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches.” Political Analysis 28 (4): 532–551. https://doi.org/10.1017/pan.2020.4.

Monarch, R. M. 2021. “Human-in-the-Loop Machine Learning.” Simon & Schuster.

Mor-Lan, G. 2019. “Transfer Learning and its Applications for Political Science Research.” Advances in Comparative Politics Workshop, Cologne Center for Comparative Politics, 2019

Mueller, H., and C. Rauh. 2018. “Reading Between the Lines: Prediction of Political Violence Using Newspaper Text.” American Political Science Review 112 (2): 358–375. https://doi.org/10.1017/S0003055417000570.

Nicholls, T. and J. Bright. 2019. “Understanding News Story Chains using Information Retrieval and Network Clustering Techniques.” Communication Methods and Measures, 13:1, 43-59, DOI: 10.1080/19312458.2018.1536972

Ophir, Y., D.K. Forde, M. Neurohr, D. Walter, and V. Massignan. 2021. “News media framing of social protests around racial tensions during the Donald Trump presidency.” Journalism, 0(0). https://doi.org/10.1177/14648849211036622

Ophir, Y., D. Walter, and E. R. Merchant. 2020. “A Collaborative Way of Knowing: Bridging Computational Communication Research and Grounded Theory Ethnography.” Journal of Communication 70 (3): 447–472. https://doi.org/10.1093/joc/jqaa013.

Pilny, A., K. McAninch, A. Slone, and K. Moore. 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.

Pustejovsky, J. and A. Stubbs. 2013. “Natural Language Annotation for Machine Learning.” O’Reilly Media, Inc.

Reimers, N., and I. Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084 [cs.CL].

Sebők, M., and Z. Kacsuk. 2021. “The Multiclass Classification of Newspaper Articles with Machine Learning: The Hybrid Binary Snowball Approach.” Political Analysis 29 (2): 236–249. https://doi.org/10.1017/pan.2020.27.

Settles, B. 2012. “Active Learning.” Synthesis Lectures on Artificial Intelligence and Machine Learning, 6(1):1–114.

Settles, B., and M. Craven. 2008. “An Analysis of Active Learning Strategies for Sequence Labeling Tasks.” In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 1070–1079.Association for Computational Linguistics. https://aclanthology.org/D08-1112.

Stoll, A., M. Ziegele, and O. Quiring. 2020. “Detecting Incivility and Impoliteness in Online Discussions.” Computational Communication Research 2, (1): 109–134. https://computationalcommunication.org/ccr/article/view/19.

Stoltenberg, D., D. Maier, and A. Waldherr. 2019. “Community detection in civil society online networks: Theoretical guide and empirical assessment.” Social Networks, 59: 120-133.

Sun, Y., M. S. Kamel and Y. Wang, 2006. "Boosting for Learning Multiple Classes with Imbalanced Class Distribution." Sixth International Conference on Data Mining (ICDM'06). doi: 10.1109/ICDM.2006.29.

Trilling, D. and M. van Hoof. 2020. “Between Article and Topic: News Events as Level of Analysis and Their Computational Identification.” Digital Journalism, 8 (10), 1317-1337, DOI: 10.1080/21670811.2020.1839352

Tsoumakas, G., and I. Katakis. 2007. “Multi-Label Classification: An Overview.” International Journal of Data Warehousing and Mining (IJDWM), 3(3), 1-13. http://doi.org/10.4018/jdwm.2007070101

Tyagi, S., and S. Mittal. 2020. “Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning.” In: Singh, P., Kar, A., Singh, Y., Kolekar, M., Tanwar, S. (eds) Proceedings of ICRIC 2019 . Lecture Notes in Electrical Engineering, vol 597. Springer, Cham. https://doi.org/10.1007/978-3-030-29407-6_17

Walter, D. and Y. Ophir. 2021. “Strategy Framing in News Coverage and Electoral Success: An Analysis of Topic Model Networks Approach.” Political Communication, 38:6, 707-730, DOI: 10.1080/10584609.2020.1858379

Weiss, G.M. 2013. “Foundations of Imbalanced Learning.” In Imbalanced Learning, edited by H. He and Y. Ma, 13-41. https://doi.org/10.1002/9781118646106.ch2

Ying, L., J. Montgomery, and B. Stewart. 2022. “Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures.” Political Analysis 30(4): 570-589. doi:10.1017/pan.2021.33

Yousefi, F., Z. Dai, C.H. Ek, and N. Lawrence. 2016. “Unsupervised Learning with Imbalanced Data via Structure Consolidation Latent Variable Model.” https://arxiv.org/abs/1607.00067

Zhao, J., M. Lan, and J. F. Tian. 2015. “ECNU: Using Traditional Similarity Measurements and Word Embedding for Semantic Textual Similarity Estimation.” In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 117–122. Association for Computational Linguistics. https://doi.org/10.18653/v1/S15-2021.

Zoizner, A., S. R. Shenhav, Y. Fogel-Dror, and T. Sheafer. 2021. “Strategy News Is Good News: How Journalistic Coverage of Politics Reduces Affective Polarization.” Political Communication 38 (5): 604–623. https://doi.org/10.1080/10584609.2020.1829762.

Appendix A.

Descriptive Statistics of Categories for Datasets used in Simulations:

20 Newgroups Dataset

The 20 Newsgroups dataset consists of nine thematic categories. In the first simulation (Graph A in Figure 6), we kept the category balance scheme 'as is', running the annotation procedure and classifier on all categories. In the second simulation (Graph B), we utilized the inherent imbalance between the categories to divide the list into large categories and small categories (below 2000 corresponding documents). In each simulation run, we would maintain the three large categories at their full size, while randomly choosing a single category from the list of small categories to undersample. This allowed us to simulate the attempted classification of a single, theory-driven, rare category.

Category	Documents	% of Corpus	Category Type
Computer	4891	25.95	Large Category
Science	2962	15.72	Large Category
Politics	2625	13.93	Large Category
Sport	1993	10.58	Small Category
Automobile	1986	10.54	Small Category
Religion	1625	8.62	Small Category
Medicine	990	5.25	Small Category
Sales	975	5.17	Small Category
Alt.atheism	799	4.24	Small Category

New York Times Front Page Dataset with CAP codes

The New York Times Front Page dataset consists of 28 macro categories, each of which consists of multiple subcategories. For our simulations, we focused only on those macro categories with at least 1000 corresponding documents (Listed below). For each simulation run, we randomly chose a single macro category to focus on, before aggregating the rest of the domains to a single category 'Other'. Then, in the simulation itself, the classifiers were tasked with classifying the multiple sub-categories within the macro domain chosen.

Macro Category	**Documents**	**% of Corpus**
International Affairs and Foreign Aid	5074	31.48
Defense	3515	21.81
Government Operations	3052	18.94
Law, Crime, and Family Issues	1794	11.13
Health	1628	10.10
Banking, Finance, and Domestic Commerce	1054	6.54

Appendix B.

Demonstrating the Conversion of Texts into Sentence Embeddings using Sentence-BERT

Sentence-BERT is one of several sentence embedding models that researchers can use. The procedure is simple – we can download a deep learning language model pre-trained on large corpora in order to represent any input text into a vector.

For example, the text “I baked the chocolate cake in the oven.” is converted into a vector with 384 dimensions (In our simulations, we utilized a model with 768 dimensions):

[6.00954443e-02 3.69346365e-02 5.25096059e-03 2.83029266e-02 -1.32560544e-03 -4.87135127e-02 -2.85588414e-03 -4.17023748e-02 -2.79687432e-04 3.44136320e-02 -2.28446592e-02
-6.32430464e-02 -6.01674654e-02 -1.22785941e-02 1.75573379e-02 -3.93791571e-02 4.12580818e-02 -1.15194723e-01 -1.75341424e-02 1.79283749e-02 1.31942183e-02 1.86974742e-
02 -1.22360764e-02 3.52652594e-02 2.31434796e-02 3.12574878e-02 3.13228220e-02 4.02686093e-03 -4.72011864e-02 -3.56128663e-02 -2.76826043e-02 -1.63963418e-02 -
2.50136806e-03 8.80754739e-03 1.66549217e-02 -4.55216644e-03 -1.88347250e-02 2.82200873e-02 4.94777448e-02 -4.93034199e-02 5.12852296e-02 2.20662318e-02 1.55101875e-02 -
1.62852947e-02 8.66450071e-02 -1.19070495e-02 1.37181515e-02 -3.92372869e-02 3.29040885e-02 4.40969169e-02 4.63366648e-03 4.23102230e-02 -1.49679743e-02 2.70957891e-02
-6.06405847e-02 -2.89411321e-02 8.90740287e-03 3.64552885e-02 4.20501344e-02 2.05845144e-02 -4.36888859e-02 3.64761427e-02 3.16999620e-03 3.18804160e-02 3.30756232e-02
-2.14112476e-02 -6.28434345e-02 2.49974001e-02 4.71758749e-03 7.89385140e-02 2.15580743e-02 6.08446561e-02 2.08110698e-02 -2.38736328e-02 -6.11377833e-03 -5.83701283e-02
1.08003333e-01 1.17955180e-02 -4.64999210e-03 4.04093191e-02 -6.04591407e-02 2.52811350e-02 7.07231015e-02 6.33231550e-02 -1.49700120e-02 4.77309786e-02 5.66647910e-02
1.17763737e-02 1.15996907e-02 -9.86885726e-02 3.79136167e-02 -4.49243337e-02 -5.83599061e-02 1.22272432e-01 -6.43654838e-02 -1.17822386e-01 -1.00646820e-02 2.27782857e-02
3.75926457e-02 6.79980144e-02 6.50774874e-03 2.77781882e-03 -2.41520721e-02 9.14280186e-04 1.58010218e-02 2.60427445e-02 1.23193357e-02 3.61967087e-02 1.08585827e-01
-2.45253518e-02 -1.50372433e-02 3.10489554e-02 1.09357707e-01 -2.56713554e-02 -6.65341541e-02 -4.39945199e-02 -6.28441423e-02 6.24561608e-02 1.06237531e-01 5.34128249e-02
3.02816276e-02 1.90563407e-02 3.54425088e-02 7.63039440e-02 -1.46215454e-01 -9.22063515e-02 2.27203220e-02 -5.07195323e-33 -6.44253893e-03 -1.93717144e-02 1.44322440e-01
7.67924562e-02 5.58511503e-02 -4.32215482e-02 -6.93855435e-02 2.56914981e-02 -1.12984218e-02 -1.67504903e-02 1.26071170e-01 6.23834785e-03 -3.51728760e-02 5.87973855e-02
-6.67829216e-02 9.18412730e-02 -7.31054544e-02 6.06388645e-03 1.07917793e-01 1.91145632e-02 -4.85623069e-02 -2.31758747e-02 -1.98296551e-02 5.31689785e-02 -4.88985777e-02
1.01617008e-01 -3.76553764e-03 -2.84355804e-02 -8.04362968e-02 3.47782448e-02 1.06035277e-01 5.91748357e-02 -8.30502342e-03 -1.69804301e-02 -2.72996575e-02 5.04218601e-02
1.29965935e-02 3.60201411e-02 -6.80425614e-02 3.41664851e-02 1.84840765e-02 -5.49914092e-02 4.67499755e-02 -3.59534547e-02 -2.97299158e-02 -3.58039252e-02 -3.09173241e-02
1.36314273e-01 1.56372935e-02 -1.13558481e-02 -4.43811379e-02 1.12404218e-02 1.11446075e-01 3.27707045e-02 -4.65956546e-04 -4.26905937e-02 6.54553951e-05 -6.83120117e-02
5.17184706e-03 1.86517984e-02 2.22640522e-02 -1.23026110e-02 -3.71525362e-02 4.33616489e-02-1.29356682e-01 -5.78531157e-03 -3.62879299e-02 2.96500307e-02 2.54503340e-02 -
6.34178147e-02 -1.08289458e-01 1.91420615e-02 9.32409763e-02 -4.02313881e-02 -2.14393269e-02 -4.19203043e-02 3.47202038e-03 1.47976102e-02 -5.92909455e-02 -3.53221036e-02
1.36448622e-01 -3.66934314e-02 -8.77170637e-02 -5.02126999e-02-2.26471126e-02 -6.36389386e-03 -7.19175041e-02 -2.18105465e-02-1.97988860e-02 7.69932792e-02 -6.66015968e-02 -
1.21422186e-02 7.61535391e-02 3.19602224e-03 -1.04757562e-01 3.61746889e-33 1.06528644e-02 -1.83819383e-02 -1.33144688e-02 1.62289720e-02-2.16021407e-02 -7.29190558e-02

7.19139576e-02 -6.60133827e-03-3.95376086e-02 2.13967133e-02 4.43418697e-02 -1.03270218e-01 3.36614400e-02 -2.75964923e-02 1.64249334e-02 2.06049830e-02 1.65559258e-02 3.19799893e-02 6.67534620e-02 -5.67634404e-02 -2.37243641e-02 5.34203015e-02 -3.88839003e-03 2.17401665e-02 -1.74921583e-02 6.39214739e-02 7.70627335e-02 1.10997155e-01 -2.39039771e-02 -5.46194538e-02 6.52641989e-03 -6.18954711e-02 -1.92693509e-02 -1.37276882e-02 -8.69988725e-02 8.49684775e-02 -1.40977520e-02 -7.96702206e-02 3.09513342e-02 -3.57431173e-02 -8.53719115e-02 1.66181475e-02 -4.05813679e-02 1.22814648e-01 9.64946076e-02 2.41774097e-02 7.45356502e-03 -3.91349085e-02 8.02878812e-02 3.99912782e-02 8.16612039e-03 -6.55903816e-02 -9.20116454e-02 -6.33838698e-02 -8.89380556e-03 5.75467050e-02 -2.37588119e-02 9.60993860e-03 2.79221237e-02 -1.13918297e-02 -5.48206940e-02 1.77656561e-02 6.04456961e-02 2.10392103e-02 -6.47352338e-02 7.03026876e-02 -2.35013235e-02 4.21596728e-02 4.43854742e-02 4.97698374e-02 -4.93135192e-02 3.28054428e-02 -6.56081783e-03 -1.77842239e-03 2.83615235e-02 -3.10261771e-02 7.29609281e-02 2.10728608e-02 -6.23399876e-02 -6.18374422e-02 -4.24383916e-02 -3.05707809e-02 7.34177381e-02 -3.29935886e-02 -1.18802143e-02 -3.66667588e-03 7.57251959e-03 -1.43914178e-01 -4.57804166e-02 -4.73074801e-03 -2.71227751e-02 6.20199926e-02 7.63770714e-02 -2.64477991e-02 1.05894804e-02 -1.55951358e-08 5.26959039e-02 -2.44627539e-02 2.71360646e-03 -2.77093537e-02 -3.18840444e-02 8.52942932e-03 4.29475158e-02 -8.43868926e-02 -5.42335846e-02 -8.60265493e-02 -4.03258651e-02 -1.02882190e-02 -9.70774051e-03 4.36043330e-02 4.18184288e-02 -9.10554975e-02 9.03847218e-02 -2.88549215e-02 -8.27697199e-03 -2.98868492e-02 -3.99242714e-02 1.01066090e-01 6.21075369e-02 -2.88961697e-02 2.52041370e-02 -3.07253301e-02 3.32841650e-02 1.43918674e-02 -4.66277413e-02 6.08059354e-02 -1.12966686e-01 4.26766835e-02 -3.99562530e-02 4.92186919e-02 -4.56726961e-02 -6.43883971e-03 -6.07347973e-02 -5.18337525e-02 -1.98892914e-02 -7.62746707e-02 9.89230070e-03 3.76399718e-02 -9.16383695e-03 -4.46242765e-02 5.00764400e-02 -7.00003328e-03 -1.53090814e-02 3.35499495e-02 3.32147861e-03 4.59059849e-02 -2.14989809e-03 5.26674949e-02 -1.62781160e-02 5.85402548e-02 1.54470745e-02 -5.98462857e-02 -5.17086452e-03 -3.38714607e-02 -4.60828422e-03 1.53943675e-03 6.43771291e-02 -3.03914938e-02 -7.97511637e-02 -9.43172872e-02]

Using such embeddings, we can now measure the semantic distances between texts. For example, the text “I baked a cake for the birthday.” is found at a cosine distance of 0.704 from the example text, and “The war in Ukraine waged, with both sides using rockets, tanks and artillery on the battlefield.” is 0.106 cosine distance from the example.

Appendix C.

Instructions for Human Coder in Second Simulation:

This simulation accompanies the manuscript "Leveraging Researcher Domain Expertise to Annotate Concepts within Imbalanced Data", and is intended to complement the computerized simulations within the paper with a demonstration of the human component of our method.

In general, the method we propose offers a procedure to improve the *annotation* stage of machine learning classification: When approaching a new, unlabelled target dataset of texts, how do we best choose which sample texts to label and include in our training set?

We will be using the 20 NewsGroups text dataset - a collection of nearly 20,000 posts from a variety of online discussion groups. This dataset is a classic collection and has been used for many machine learning classification tasks.

We will be working with 3 of the categories within the 20 NewsGroup collection:

Sports - Baseball, **Sports - Hockey**, and **Politics - Middle East**. The rest of the categories will be aggregated into a single **Other** category.

We provide here a couple of sample posts for each category of interest:

Baseball:

5. *Would someone please give me the address for Texas Ranger*

   ticket orders. Thanks very much.

6. *: Hi there,: I'm german and I have been into this MLB stuff since almost one year now. : There are many problems occuring for me. One of them is the ERA statistic for : pitchers. What does it say ??*

ERA indicates the average number of earned runs attributed to a pitcher per

nine inning game. Thus, if a pitcher pitched 3 innings and gave up 1 earned

run, his 9 inning equivelent perfomance would be 3 earned runs, thus his ERA

is 3.00. To compute the ERA you simply take the number of earned runs divided by the innings pitched and then multiple the result by 9.

*ERA = (ER/IP) * 9*

An earned run is run that is given up by the pitcher that is not attributed to a fielding error. More specifically, if an error occurs that represented the third out, all runs scored after the error are considered UNEARNED runs. Earned runs are also runs scored as a result of players who were left on base when the pitcher exited the game.

Hockey:

7.

Ottawa picks first, because they had fewer wins during the season, the first tiebreaker.

--

Keith Keller *LET'S GO RANGERS!!!!!*

LET'S GO QUAKERS!!!!!

kkeller@mail.sas.upenn.edu *IVY LEAGUE CHAMPS!!!!*

8. Ok, but have you seen Tabaracci play yet? In his two starts and his relief effort for Beaupre, he has looked mighty sharp - don't forget the shutout. I think he's let in just four goals over eight periods of play. I like Hrivnak, but we might actually have to give some credit to David Poile for a change after this trade.

Hopefully if Tabaracci starts against the Isles tonight, I haven't jinxed him.

Middle East Politics:

9. This is historically incorrect. Early Zionist 'fighters' did indeed target civilians. They made random attacks in Arab marketplaces, killing innocent passers-by. Your assertion of the opposite is an

attempt to whitewash history. Anyone can read about the history of the Zionist terrorists. A good book to start is the one by J. Bowyer Bell, an expert in international terrorism. (His main interest is Irish terrorism.)

10. As we see right now, the position of influence enjoyed by parties favoring the negotiation process is tenuous at best. The local "elections" in Hebron that the PLO was expected to win (perhaps adding a bit to its flagging position of "legitimacy" in the eyes of Palestinians and the Middle East) have been disrupted by Hamas actions overtly directed towards undermining those (and all West Bank) elections. The present ruling Israeli Labor coalition seems to be one rather thin political ice. The Palestinian delegation has been reduced from 14 to three to protest Israel "lack of seriousness" in the talks and refusal to reverse all the deportations immediately.

Instructions for Human Coder:

Stage 1 of Simulation - Creating core collections of posts for each category of interest

At this stage, we need you to create a collection of 10 posts that are relevant for each of the three categories (40 total). Please try to capture a wide, diverse representation for each category - instead of using a single keyword to collect extremely similar posts, try to find original or different posts that belong to the same category.

In the "20NewsGroups_textsonly.csv" file, we have assembled the full corpus of texts, with their labels removed. Using this collection, you can search for and target posts that you deem to belong to each one of the categories.

Please record the procedure used to target texts. For example, write down what keywords you used to search for relevant posts (you can choose any strategy that seems appropriate, using single or multiple keywords together), and note the number of texts read before choosing the 10.

Regarding the "Other" category:

Note - any text you read that does not fall within the target category, please label as "Other" and save. At the end of this stage we will want 30 labeled texts for each category, before making sure to have an additional 10 posts for the "Other" category. If you have reached 10 "Other" posts simply through the searches for the previous 30 posts, this is sufficient. If, however, you still

have less than 10 “Other” posts, simply take a random sample of the missing number of posts from the full csv file and label correspondingly.

We realize that this stage is the most critical, yet perhaps ambiguous of the simulation. Thus, if you have read 50 posts in total, yet are still haven’t reached 10 core texts for any of the four categories, please let us know and we will decide how best to proceed.

At this stage, note the indices of each of the posts collected as relevant for each category and report back to us (together with your procedure used).

Stage 2 - Manual Coding of Posts

At this stage, we will begin to run our method of collecting additional texts to be added to our expanding training set.

We will provide you at each round (between 2-3 rounds, this will be determined as we continue the simulation) with a list of new texts, unlabeled. Your task is to label each post with the label you deem to be most appropriate: **Baseball**, **Hockey**, **Middle East,** or **Other** (if none of the categories of interest are present in the text.

You will return these labels to us and we will continue to run through the rounds of our method.

Additional Points

- While most posts might be easily recognizable, we note that many of them make references to players, actors, sports teams, etc. that you may not be familiar with. It is ok to look up any such names or references in order to determine the category.
- Please note at each stage the number of documents read (especially the first) as this is the measure we are attempting to track throughout this simulation.

Feel free to contact us with any questions at any stage of the process!

Appendix D.

Posts Selected for Human Labeling with Given Labels:

In the attached spreadsheet, we have included the full training set of posts from the 20 Newsgroups dataset, that have received labels by our human coder in the second simulation. Such posts were collected in three rounds:

1) Our human coder was tasked with finding 10 representative posts for each category, using keywords to filter the full corpus.
2) First round of stratified sampling for each category on the shared semantic space. Samples were then passed on to the human coder for labeling, while being oblivious to the semantic distance of each post from each centroid.
3) Second round of stratified sampling with each centroid having been updated to include the posts from the previous round.

Spreadsheet found at: https://dataverse.harvard.edu/privateurl.xhtml?token=ec8787c1-1d3a-4f0c-b999-0cc21adaf4f4