# Active Learning for Corpus Refinement: Cost-Effective Preprocessing to Improve Validity of Applied Quantitative Text Analysis

**Jakob Steglich, Stephan Poppe**
Institute of Sociology
Leipzig University
{jakob.steglich,stephan.poppe}@uni-leipzig.de

## Abstract

Quantitative text analysis relies on high-quality corpora, but keyword-based collection often retrieves irrelevant material, undermining validity. We show that active learning with a transformer-based classifier can iteratively refine corpora by excluding irrelevant documents, prompting researchers to clarify inclusion criteria and address edge cases. Applied to German newspaper articles on depression and schizophrenia, this approach improves construct validity and reduces labeling effort. The document relevance classifiers reached an F1-score of 0.8 with just 100–150 labeled snippets, with further gains from tuning, outperforming both random sampling and a weakly supervised sampling baseline. Filtering non-medical articles further had little effect on downstream depression stigmatization measures but increased schizophrenia stigmatization. Active learning thus enables efficient corpus validation and clearer concept boundaries with minimal preprocessing. The source code is publicly available at https://github.com/jakobstgl/active-learning-corpus-refinement.

## 1 Introduction

Recent years have seen a rapid increase in automated text analyses in the social sciences (Stoltz and Taylor, 2024; Grimmer et al., 2022). Researchers rely on large text corpora to extract meaning, for example, through sentiment analysis. Beyond dictionary-based approaches, machine learning techniques such as topic modeling (DiMaggio et al., 2013), or word-embedding-based relation extraction (Stoltz and Taylor, 2021; Arseniev-Koehler and Foster, 2022; Kozlowski et al., 2019; Nelson, 2021; Boutyline and Arseniev-Koehler, 2025), have become standard tools in computational social science research.

However, the construction of the underlying corpus remains a critical challenge. Text data are typically retrieved via keyword searches based on surface forms rather than semantic relevance (Grimmer et al., 2022). As a result, these methods often fail to capture only those documents that reflect the construct defined by the research question (Hanny et al., 2024; Hanani et al., 2001). Therefore, corpora frequently include conceptually irrelevant material, which can introduce systematic bias and compromise the validity of downstream analyses (Grimmer et al., 2022, p. 41-46).

A core reason for this problem lies in lexical ambiguity. In fact, in natural language processing (NLP), polysemy is a long-standing issue (Bevilacqua et al., 2021), particularly when corpora are assembled via keyword search. For example, in media reporting on mental illness, the term "depression" may refer either to a medical condition or to an economic downturn. Similarly, "schizophrenia" is frequently used metaphorically to describe contradictory or incoherent situations.

Research on word-sense disambiguation (WSD) has proposed various solutions to these problems (Bevilacqua et al., 2021), including rule-based approaches and supervised classifiers trained on annotated data (Banerjee and Pedersen, 2002; Viveros-Jiménez et al., 2013; Kapitanov et al., 2019). While effective, these methods are typically resource-intensive and are often applied as separate preprocessing steps rather than being integrated into the main analytical pipeline. Recent work has begun to use active learning (Lewis and Gale, 1994) for WSD (e.g. Wang et al., 2018; Zhu and Hovy, 2007), but this research focuses mostly on benchmark performance and does not assess the downstream impact on relevant outcomes of analyses.

In this paper, we therefore address the following research questions:

1. Can document relevance classification for socio-psychological constructs be improved with uncertainty-based active learning?
2. What is the effect of higher corpus quality



*Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*
*Volume 4: Student Research Workshop*, pages 952–966

on downstream analyses of these constructs, specifically the stigmatization of depression and schizophrenia?

First, we establish an active-learning-based document classification approach that integrates corpus refinement directly into the research pipeline. We apply this approach to a corpus of German newspaper articles on depression and schizophrenia. Both illnesses occur in medical and non-medical senses. However, for our downstream analysis, we are only interested in the medical usage. Our method to iteratively exclude irrelevant documents from the corpus leverages transformer-based models and results in improved construct validity with minimal labeling effort.

Second, we introduce a downstream task that examines the stigmatization of these diseases. Our methodology is adapted from Best and Arseniev-Koehler (2023) and operationalizes stigmatization as the proximity of disease embeddings extracted from newspaper articles to a latent semantic dimension in the embedding space. This dimension is constructed using anchor embeddings of stigmatizing and anti-stigmatizing terms. For instance, psychiatric disorders are often associated with negative personality traits (e.g., depression being framed as laziness). A corresponding stigma dimension is defined as the difference between embeddings of negative and positive character traits. Disease embeddings located closer to the negative pole of this dimension are interpreted as more stigmatized in the corpus, with cosine similarity serving as the stigma score. We then examine whether the distribution of stigma scores for depression and schizophrenia shifts after removing irrelevant documents. The full analysis pipeline is displayed in Figure 1.

Our approach further demonstrates how active learning can support corpus preprocessing by encouraging researchers to define clear inclusion and exclusion criteria to resolve edge cases that challenge the boundaries of the research construct.

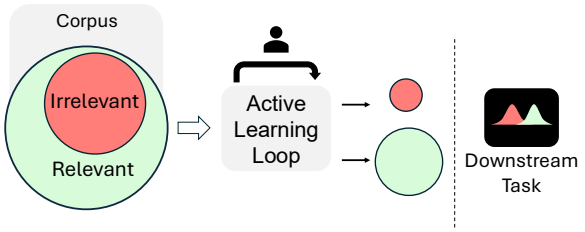


Figure 1: Analysis Pipeline. (1) Corpus Acquisition: Gathering relevant and irrelevant documents. (2) Iterative Refinement: Active learning cycles where experts label boundary cases to improve classifier robustness. (3) Global Prediction: Classification of the full corpus based on learned relevance. (4) Downstream Validation: Assessing the shift in metric distributions resulting from the removal of irrelevant data.

## 2 Related Works

### 2.1 Active Learning in NLP

Active learning is a machine learning paradigm introduced by Lewis and Gale (1994), in which a model is iteratively retrained on selectively labeled data. The process typically begins with a small labeled dataset used to train an initial classifier. In each subsequent iteration, a query strategy selects the most informative instances from a pool of unlabeled data. These instances are labeled by a human annotator (the oracle), added to the training set, and used to update the classifier. This loop continues until a stopping criterion is met, such as performance convergence or the exhaustion of a labeling budget.

The choice of a query strategy is central to the success of active learning, as it determines which instances are selected for annotation (Kumar and Gupta, 2020). Uncertainty-based strategies, for example, prioritize instances for which the model's predictions are least confident. Labeling such high-uncertainty samples has been shown to accelerate learning and yield higher performance compared to random sampling (Miller et al., 2020; Jacobs et al., 2022). While some strategies rely on prediction confidence (such as the prediction entropy strategy), others exploit information from embedding distances or gradient-based criteria (Schröder et al., 2023, 2022). In a comprehensive evaluation of active learning strategies for transformer-based text classification, Schröder et al. (2022) show that the breaking ties strategy is most effective. In a binary classification task, breaking ties is equivalent to the prediction entropy strategy (Roy and McCallum, 2001).

In addition, numerous approaches have been proposed to further extend active learning and improve performance (Margatina et al., 2021; Korakakis et al., 2024; Zhang et al., 2022; Deng et al., 2023). However, as our priority is to investigate the effectiveness of active learning for corpus refinement

and to assess the downstream effects of such refinement, the incorporation of more advanced active learning strategies is left for future work.

### 2.2 Active Learning for Word-Sense Disambiguation

Previous research has applied active learning and related paradigms to problems such as word-sense disambiguation (WSD) and corpus filtering. For example, Wang et al. (2018) demonstrate that an interactive learning algorithm, closely related to active learning, can substantially reduce labeling effort when distinguishing ambiguous meanings of medical terms in WSD datasets. More generally, the introduction of transformer-based language models (Vaswani et al., 2017) has improved the effectiveness of active learning frameworks (Jacobs et al., 2022), including applications to corpus refinement. For instance, Hanny et al. (2024) show that a RoBERTa-based (Liu et al., 2019) active learning approach outperforms alternative strategies with minimal labeling efforts when identifying disaster-related tweets. Despite these advances, existing work primarily evaluates active learning approaches on benchmark tasks. What remains largely unaddressed is the extent to which corpus refinement through active learning affects the substantive results of downstream analyses.

### 2.3 Stigma in Newspaper Reports

Stigmatization is a social process in which human differences are labeled and subsequently linked to negative stereotypes, such as negative personality traits (Link and Phelan, 2001; Phelan et al., 2008). These stereotypes often form the basis for discrimination against affected individuals. One theorized mechanism behind the formation of such stereotypes is the social enforcement of norm conformity: individuals who deviate from perceived norms are labeled as outsiders and stigmatized in an effort to maintain group cohesion (Phelan et al., 2008; Link and Phelan, 2014). Many survey studies find prevailing stigmatizing attitudes toward people with mental illness, with notable variation across different diagnoses (Schomerus et al., 2022; Pescosolido et al., 2021). Media coverage also plays a key role in the reproduction of such stigmas. Qualitative research has shown that news articles provide culturally available frames of stigmatization, for instance by associating those suffering from schizophrenia with dangerousness and unpredictability (Corrigan et al., 2004; Sittner et al., 2024). These representations shape public perceptions, as readers internalize media narratives and project these stigmas onto others (Best and Arseniev-Koehler, 2023).

## 3 Methods: Active Learning for Corpus Validation

Beyond its computational efficiency, we argue that active learning offers particular advantages for social research, where research topics are often theoretical constructs whose operationalization is not self-evident. This also applies to seemingly well-defined concepts such as diseases. While standardized classifications like the International Classification of Diseases (ICD) (World Health Organization, 2019/2021) provide clear clinical definitions, their usage in natural language is considerably more ambiguous. In textual corpora, disease terms occur in borderline, colloquial, or metaphorical contexts, making relevance decisions non-trivial and dependent on theory.

Moreover, even before determining the contextual framing of such references, researchers must make deliberate methodological decisions on the types of usage they aim to investigate. Both the decision about whether a document is relevant for the research question, as well as the degree to which the concept is prevalent in the article, thus depend on the subjective expertise of the researcher. Active learning is especially suited here for two reasons. First, the learning loop explicitly returns samples for which the prediction is uncertain (Miller et al., 2020). These samples not only help the model generalize better, but also challenge the researcher in the definition of the scope of their research, since they have to decide on edge cases. Secondly, in contrast to a zero-shot learning approach, active learning enables experts to provide domain knowledge. They can thus actively partake in how the classifier learns a certain concept of interest (Wang et al., 2018; Miller et al., 2020).

### 3.1 Analytical Strategy

Our analytical strategy comprises two main steps. First, we compute a stigma score for each document in a large corpus of German newspaper snippets on depression and schizophrenia, following the embedding-based method proposed by Best and Arseniev-Koehler (2023). While stigmatization is not the primary object of investigation, these scores serve as a downstream outcome with which we evaluate the effects of our corpus refinement

procedure.

Second, an iteratively trained active-learning-based classifier distinguishes medically relevant references to mental illness from non-medical or metaphorical uses. After training, we assess whether the filtered corpus differs meaningfully from the full corpus with respect to stigma score distributions. Specifically, we compare the distributions visually and quantify differences using effect sizes. All analyses are conducted separately for depression and schizophrenia to allow for diagnostic contrasts.

### 3.2 Criteria of Inclusion and Exclusion

Before initiating the active learning loop, we defined explicit inclusion and exclusion criteria to guide labeling decisions. Our objective was to retain only those documents in which mental illness terms are used in a medical and colloquial mental health context. This includes snippets describing symptoms, diagnosis or treatment, personal accounts, or societal impact of depression and schizophrenia. In contrast, we exclude metaphorical uses (e.g. "economic depression") as well as other polysemous meanings that are unrelated to mental health.

## 4 Experimental Setup

### 4.1 Dataset

We use data from the Mannheim German Reference Corpus (DeReKo), a large archive of German-language texts (Kupietz et al., 2010). Licensing restrictions limit access to short text snippets. In our corpus, these snippets have an average length of 70 words.

We retrieved documents using keyword-based searches. The initial keyword lists for depression and schizophrenia were taken from Best and Arseniev-Koehler (2023) and extended using terminology from the International Classification of Diseases (ICD-11) (World Health Organization, 2019/2021). Data collection was automated using `Selenium` (Gojare et al., 2015). In total, we collected 631,176 newspaper snippets published between 2000 and 2024. Of these, 516,382 snippets contain keywords related to depression and 114,794 to schizophrenia. As a result of the keyword-based retrieval strategy, the raw corpus inevitably contains a substantial number of non-medical and metaphorical uses of disease terms, motivating the corpus refinement approach described below.

### 4.2 Corpus Filtering Evaluation

#### 4.2.1 Active Learning Procedure

We implement active learning using the `small-text` library in Python (Schröder et al., 2023), which provides a modular framework for combining query strategies with transformer-based text classifiers. Since all classification tasks in our study are binary, we employ a prediction-entropy query strategy (Roy and McCallum, 2001), which prioritizes instances for which the model exhibits the highest class uncertainty.

We trained two classifiers using active learning, one for depression and one for schizophrenia, following the same learning protocol. For each task, we randomly sampled 100 instances as a fixed test set and 50 instances as an initial labeled training set. In each active learning iteration, we annotated batches of 25-100 instances, depending on the model performance (Figure 3). After labeling 350 examples per classifier, we halted training due to self-imposed budget constraints.

#### 4.2.2 SetFit for Classification

As classifier, we employ SetFit (Tunstall et al., 2022), an approach based on a transformer-based encoder model (Vaswani et al., 2017) optimized for a few-shot learning setup. SetFit combines embeddings from a pre-trained sentence-encoder (Reimers and Gurevych, 2019) with a contrastive step, followed by classification using logistic regression. During contrastive training, the model constructs positive and negative pairs of sentences based on class labels and adjusts the embedding space to bring semantically near pairs closer together while pushing dissimilar ones apart (Tunstall et al., 2022). This procedure enables efficient learning from small labeled datasets. In our case, after contrastive training of the embedding space, a simple classifier predicts whether a snippet refers to a mental illness in a medical context or not. We choose the *paraphrase-multilingual-MiniLM-L12-v2* Sentence Transformer model as a backbone because of its computational efficiency and robust multilingual support (Reimers and Gurevych, 2019).

#### 4.2.3 Baselines

We compare the proposed active learning approach against two baseline sampling strategies. All meth-

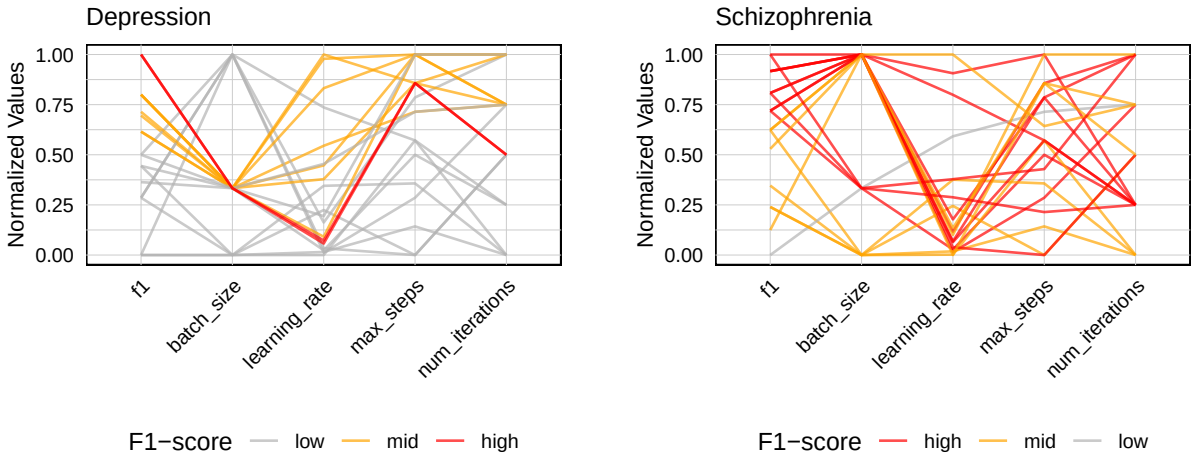


Figure 2: Hyperparameter optimization for the depression (left) and schizophrenia (right) model. High = F1 > 0.8, mid = F1 > 0.75, low = F1 <= 0.75. Each line represents a separate model that was trained.

ods start from the same labeled seed set as the active learner and query three batches of 100 samples each. As a first baseline, we employ a random sampling strategy which selects instances uniformly at random for annotation. Second, as a weakly supervised baseline, we implement an iterative bootstrapping procedure based on cosine similarity. Sentence embeddings, extracted from the *paraphrase-multilingual-MiniLM-L12-v2* model (Reimers and Gurevych, 2019), serve as the foundation. At each iteration, cosine similarities between all unlabeled instances and the current labeled set are calculated separately for each class. Each unlabeled instance is then assigned a pseudo-label corresponding to the class with the highest maximum similarity. Confidence is defined as the margin between maximum similarity scores for the two classes. The top-$N$ most confident instances are added to the labeled set, removed from the unlabeled pool, and the process is repeated until the labeled set size reaches the same size as in the active learning setup.

#### 4.2.4 Model Evaluation

Given the significant class imbalance, where *'relevant'* instances constitute the majority, we report the F1-score as the primary evaluation metric. We focus on the F1-score for the minority class (irrelevant), because we are primarily interested in the identification of such sparse samples. To provide a holistic view of the classifier's performance across both classes and prevent our classifier from labeling *'irrelevant'* too aggressively, we report the F1-macro as well.

#### 4.2.5 Training Parameter Optimization

Initially, we trained all models using the default SetFit hyperparameters: a batch size of 16, a learning rate of $2 \times 10^{-5}$, and a single training epoch. To further improve classification performance, we conducted hyperparameter optimization using the `Optuna` library in Python (Akiba et al., 2019). This optimization process is carried out using the full 350 samples from the uncertainty based active learning method for each disease. We evaluated 20 hyperparameter configurations by varying the learning rate ($10^{-6}$ to $10^{-4}$), batch size (16–32), the maximum number of training steps (20–300) and the number of iterations used to generate sentence pairs (10-50). The resulting classifiers were subsequently employed to generate inclusion and exclusion predictions for the entire corpus.

### 4.3 Downstream Task

#### 4.3.1 Further Preprocessing Steps

For the downstream task, we applied further preprocessing steps. We excluded all snippets containing fewer than 20 words and identified near-duplicate snippets published in the same year by calculating cosine similarity scores between tf-idf representations. Documents with a similarity greater than 0.85 are reduced to one instance. After preprocessing, the final dataset comprised 507,440 snippets, of which 427,078 were related to depression and 80,362 to schizophrenia. In contrast to the active learning analysis, stopwords and punctuation were removed for the computation of stigma scores. Finally, all lexical variants of disease terms (e.g. "depressed" and "depressive") were normalized to a

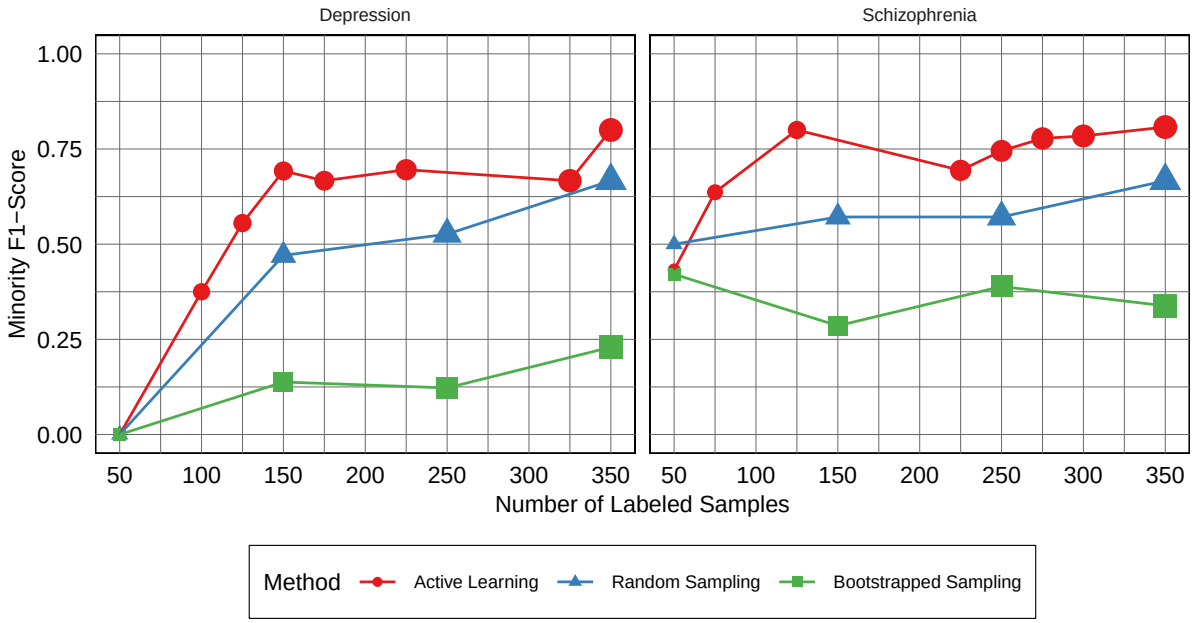


Figure 3: Evolution of the F1-score for the minority class on the test-set by number of labeled examples and sampling method for schizophrenia and depression.

single keyword (e.g. "depression"). These steps reflect the requirements for the static word embeddings on which the downstream task is based.

#### 4.3.2 Measuring Stigma

To measure stigma in our corpus, we first define a latent semantic dimension that captures stigmatization using word embeddings. Here, stigmatization reflects how strongly a disease is reported about in the context of negative personality traits. To construct the stigma dimension, we compute average embeddings for two sets of character trait keywords, one representing positive traits and one representing negative traits. The vector difference between these two averages yields a vector which is often referred to as a semantic direction in the embedding space (Arseniev-Koehler and Foster, 2022; Stoltz et al., 2024).

Second, we project each document into this dimension to obtain a stigma score. Each snippet's stigma score is computed as the cosine similarity between the embedding of the disease term and the stigma dimension. To improve interpretation, we z-standardize these scores with respect to the distribution of cosine similarities of all other words appearing in the corpus. A score of 1 indicates that the disease term in a given snippet is located one standard deviation closer to the negative traits (stigmatizing) pole of the dimension than the average of all other words in the corpus (Best and Arseniev-Koehler, 2023).

Because our data consists of short newspaper snippets, we adopt the à la carte (ALC) embedding approach (Khodak et al., 2018; Rodriguez et al., 2023b) implemented in the `ConText` package in `R` (Rodriguez et al., 2023a), which adapts pre-trained word vectors to the local context of a smaller corpus. As pre-trained embeddings, we use FastText embeddings (Bojanowski et al., 2017) trained on the German Wikipedia (Wirsching et al., 2025) with a context window of 10.

## 5 Results

### 5.1 Filtering Performance

Figure 3 shows classification performance as a function of labeled instances for active learning and two baselines, separately for depression and schizophrenia. Both active learning models improve rapidly within the first 120–150 labeled instances, although the schizophrenia model learns the task more efficiently. With only 50 labeled instances, the schizophrenia classifier already reaches an F1-score of 0.44, whereas the depression model fails to generalize at this stage due to severe class imbalance in the initial sample. Despite differing query sizes, learning trajectories are similar. The depression model benefits strongly from the initial large query, while later iterations yield diminishing returns. After annotating 350 instances, we termi-

| | Depression | | | Schizophrenia | | |
|---|---|---|---|---|---|---|
| Subset | Mean ± SD | $N$ | % | Mean ± SD | $N$ | % |
| Unfiltered | 1.33 ± 1.22 | 427,078 | 100 | 1.80 ± 1.35 | 80,362 | 100 |
| Medical | 1.40 ± 1.21 | 370,729 | 87 | 2.10 ± 1.30 | 60,374 | 75 |
| Non-Medical | 0.84 ± 1.12 | 56,349 | 13 | 0.89 ± 1.09 | 19,988 | 25 |

Table 1: Mean and standard deviation of the stigma score distribution for depression and schizophrenia by the subset of snippets.

nated the active learning loop, achieving F1-scores of 0.80 (depression) and 0.81 (schizophrenia), although strong performance was already reached after three iterations (100–120 instances). This indicates that the learning loop could have been stopped earlier. Across all iterations, active learning consistently outperforms both baselines. Random sampling reaches final F1-scores of 0.67 for both diseases, while cosine-similarity bootstrapping performs substantially worse (0.22 for depression, 0.32 for schizophrenia).

As shown in Figure A.2, the F1-macro exhibits a similar performance trajectory to Minority F1 but yields consistently higher absolute scores, while simultaneously narrowing the performance margins between sampling methods (see Tables A.1, A.2, A.3 for exact figures). Especially in later sampling iterations, active learning and random sampling show more similar results, though the superiority of active learning in the remains visible, especially in the early iterations.

To qualitatively illustrate the corpus filtering results, examples of snippets with the highest predicted probability of belonging to the *non-medical* class include:

> "This bold move is intended to further increase the money supply and stabilize the financial system in the midst of the worst economic crisis since the Great Depression of the 1930s."

> "[...] we have a schizophrenic situation. Bremen is actually a rich state; we are among the leaders in terms of millionaire income, yet the public coffers are empty [...]."

These examples are clearly irrelevant to our study of mental health conditions in a clinical context and therefore validate the quantitative results of our approach.

#### 5.1.1 Hyperparameter Tuning

To evaluate the hyperparameter tuning results for the active learning models, we visualize all tested configurations in Figure 2. For the depression classifier, high-performing configurations consistently use a batch size of 16, low learning rates ($6 \times 10^{-6}$ and $8 \times 10^{-6}$), and larger numbers of training steps. The best configuration achieves an F1-score of 0.83, improving performance by 0.03 over the best model obtained after active learning. For schizophrenia, the results are less consistent, with strong configurations spread across the parameter space and no single dominant setting beyond batch sizes of 16 or 32. Nevertheless, the best model again reaches an F1-score of 0.83. The exact hyperparameters of the best models are reported in Table A.4.

### 5.2 Impact on Downstream Stigmatization Analysis

Figure 4 displays the kernel density estimates of stigma scores before and after filtering. For depression, the mean stigma score increases only marginally from 1.33 in the unfiltered corpus to 1.40 in the medical-only subset, indicating that filtering has little impact on the aggregate stigmatization. In contrast, the distribution for schizophrenia showcases a more substantial shift: the mean score increases from 1.80 to 2.10, suggesting substantially stronger stigmatization once non-medical uses are removed. Consistent with this interpretation, the non-medical subsets of both corpora exhibit lower and more similar average scores (0.84 for depression and 0.89 for schizophrenia). This pattern indicates that the filtering procedure successfully separates semantically distinct types of usage that would otherwise mislead the downstream analysis. As reported in Table 1, these distributional differences also reflect the relative sizes of the filtered subsets. For depression, only 13% of the corpus is excluded as non-medical, whereas 26% of schizophrenia-related articles are removed.

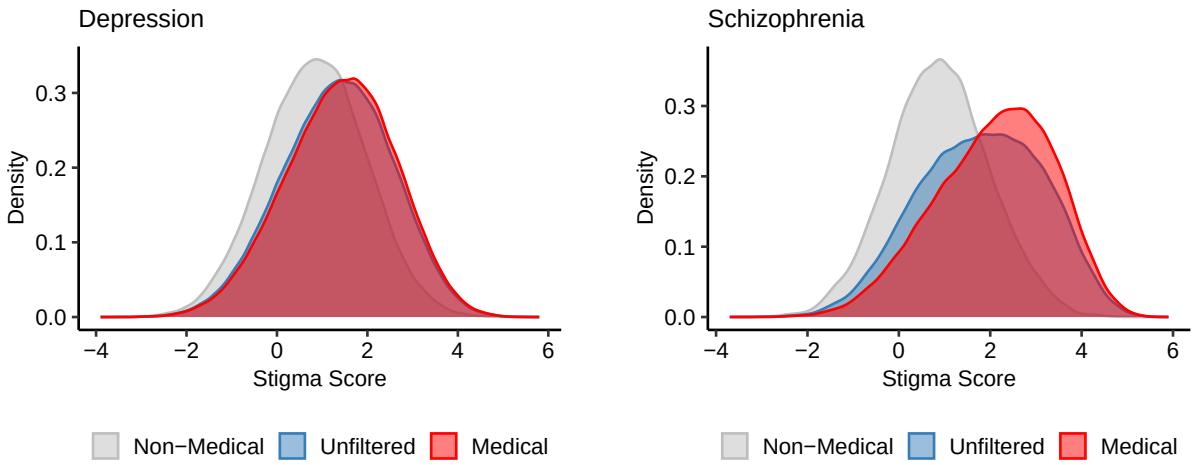


Figure 4: Kernel Density Estimate (KDE) plots for the distributions of stigma scores by disease.

This larger proportion of non-medical usage likely amplifies the effect of filtering in the schizophrenia corpus.

#### 5.2.1 Comparison of Effect Sizes

To quantify distributional differences, we compute pairwise effect sizes using Cohen's $d$ (Figure A.1). Comparing medical-only subsets with the unfiltered corpora shows a small effect for depression ($d = 0.06$) and a moderate effect for schizophrenia ($d = 0.23$). We further compare stigma score distributions between depression and schizophrenia before and after filtering, a contrast central to mental health research (Kilian et al., 2021). Filtering reinforces this difference: the effect size increases by 0.20 when only medically relevant articles are retained, indicating a more pronounced stigmatization gap once non-medical uses are removed.

## 6 Discussion

Our analyses demonstrate the importance of systematic corpus preprocessing for applied text analysis. In the case of depression, substantially similar results could be achieved even without excluding irrelevant documents. However, this appears to be more a matter of coincidence than validity. By contrast, the schizophrenia corpus illustrates that filtering can substantially alter the distribution of stigma scores. Together, these findings suggest that while corpus refinement may not always be strictly necessary to obtain stable results, it constitutes a sufficient method for construct validity, as its importance cannot be determined in advance. In addition, the human-machine interaction inherent in active learning proves particularly valuable for clarifying the scope of the research construct. The human-in-the-loop setup supports the iterative refinement of an extensional decision boundary, as encoded in the classifier, while simultaneously requiring researchers to make principled, intensional decisions about inclusion and exclusion when labeling ambiguous cases.

In contrast to passive and weakly supervised alternatives, the proposed active learning approach achieved robust classification performance with relatively little labeled data. Across all iterations, active learning consistently outperforms both random sampling and cosine-similarity-based bootstrapping on F1-minority and F1-macro, demonstrating that its gains are not merely driven by additional supervision but by the targeted selection of informative instances. These findings are in line with prior work highlighting the efficiency of active learning (Jacobs et al., 2022; Hanny et al., 2024; Miller et al., 2020). Additional performance gains from hyperparameter tuning were modest.

More broadly, our findings support the view that corpus validation and conceptual sharpening should be treated as integral components of the analytical pipeline. For example, the static embedding-based stigma scoring approach by Best and Arseniev-Koehler (2023) could be extended into a supervised classification task, where active learning is used to jointly classify document relevance and stigmatization in a multiclass framework.

We see multiple paths for future improvement. Explainability frameworks such as `Captum` (Kokhlikyan et al., 2020) could help to identify the semantic features driving the classifier's decisions,

thereby supporting a more intensionally constituted definition of the research concept.

Finally, we deliberately decided not to use large language models (LLMs) for annotation in order to retain control over the labeling process and avoid the propagation of model-internal biases (Abid et al., 2021; Naous et al., 2024). Nevertheless, hybrid annotation schemes in which LLMs act as auxiliary raters within the active learning loop could improve reliability and help uncover systematic bias in both human and machine judgment.

## 7 Conclusion

This paper underscores the importance of systematic corpus preprocessing in quantitative text analysis, showing that the inclusion or exclusion of irrelevant documents can substantially affect downstream results and substantive conclusions. Moreover, active learning enables rapid improvement in classification performance with minimal labeled data while integrating corpus refinement and concept validation through an iterative, human-in-the-loop workflow.

## Limitations

Our study has several limitations. First, the test sets used to evaluate classification performance consisted of only 100 randomly drawn snippets per condition. These samples may not fully capture the diversity of the underlying corpus, which limits the precision and generalizability of the reported performance estimates. Furthermore, we do not report confidence intervals for the classification results, as our labeling budget did not allow us to repeat the labeling process multiple times. Consequently, we cannot entirely rule out the possibility that the observed performance gains from active learning are attributable to random variation.

Second, the labeling process was conducted by a single annotator rather than multiple independent coders. Although the annotator is a graduate student with experience in the field of mental illness stigmatization research, we cannot assess inter-coder reliability. This introduces the risk of systematic labeling bias. In particular, it cannot be ruled out that some inclusion decisions reflect implicit assumptions about stigmatization rather than strictly non-medical ones. Since these decisions directly propagate into the downstream analysis, such potential biases may have affected the results. Future work should therefore incorporate multiple annotators to evaluate the labeling regime more transparently.

Third, the embedding-based stigma scoring used as the downstream task represents only one particular class of applications in computational social science. It remains unclear to what extent the observed effects of corpus refinement generalize to other downstream tasks, such as transformer-based text classification or topic modeling.

Finally, although we compare active learning against random sampling and a weakly supervised bootstrapping approach, we do not evaluate its performance relative to more specialized corpus filtering or word-sense disambiguation methods. A more comprehensive comparison across alternative refinement strategies would further strengthen the empirical assessment of active learning's advantages for corpus validation and downstream inference.

## References


Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2623–2631, New York, NY, USA. Association for Computing Machinery.

Alina Arseniev-Koehler and Jacob G. Foster. 2022. Machine Learning as a Model for Cultural Learning: Teaching an Algorithm What it Means to be Fat. *Sociological Methods & Research*, 51(4):1484–1539.

Satanjeev Banerjee and Ted Pedersen. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *Computational Linguistics and Intelligent Text Processing*, pages 136–145, Berlin, Heidelberg. Springer.

Rachel Kahn Best and Alina Arseniev-Koehler. 2023. The Stigma of Diseases: Unequal Burden, Uneven Decline. *American Sociological Review*, 88(5):938–969.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: a survey. In *Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146. Place: Cambridge, MA Publisher: MIT Press.

Andrei Boutyline and Alina Arseniev-Koehler. 2025. Meaning in hyperspace: Word embeddings as tools for cultural measurement. *Annual Review of Sociology*, 51.

Patrick W. Corrigan, Fred E. Markowitz, and Amy C. Watson. 2004. Structural Levels of Mental Illness Stigma and Discrimination. *Schizophrenia Bulletin*, 30(3):481–491.

Xun Deng, Wenjie Wang, Fuli Feng, Hanwang Zhang, Xiangnan He, and Yong Liao. 2023. Counterfactual Active Learning for Out-of-Distribution Generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11362–11377, Toronto, Canada. Association for Computational Linguistics.

Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6):570–606.

Satish Gojare, Rahul Joshi, and Dhanashree Gaigaware. 2015. Analysis and Design of Selenium WebDriver Automation Testing Framework. *Procedia Computer Science*, 50:341–346.

Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as data: a new framework for machine learning and the social sciences*. Princeton University Press, Princeton. OCLC: on1295105650.

Uri Hanani, Bracha Shapira, and Peretz Shoval. 2001. Information filtering: Overview of issues, research and systems. *User modeling and user-adapted interaction*, 11(3):203–259.

David Hanny, Sebastian Schmidt, and Bernd Resch. 2024. Active Learning for Identifying Disaster-Related Tweets: A Comparison with Keyword Filtering and Generic Fine-Tuning. In *Intelligent Systems and Applications*, pages 126–142, Cham. Springer Nature Switzerland.

Pieter Floris Jacobs, Gideon Maillette de Buy Wenniger, Marco Wiering, and Lambert Schomaker. 2022. Active Learning for Reducing Labeling Effort in Text Classification Tasks. In *Artificial Intelligence and Machine Learning*, pages 3–29, Cham. Springer International Publishing.

Andrey Kapitanov, Ilona Kapitanova, Vladimir Troyanovskiy, Vladimir Ilyushechkin, and Ekaterina Dorogova. 2019. Clustering of Word Contexts as a Method of Eliminating Polysemy of Words. In *2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pages 1861–1864. ISSN: 2376-6565.

Mikhail Khodak, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora. 2018. A La Carte Embedding: Cheap but Effective Induction of Semantic Feature Vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Melbourne, Australia. Association for Computational Linguistics.

Carolin Kilian, Jakob Manthey, Sinclair Carr, Franz Hanschmidt, Jürgen Rehm, Sven Speerforck, and Georg Schomerus. 2021. Stigmatization of people with alcohol use disorders: An updated systematic review of population studies. *Alcoholism, Clinical and Experimental Research*, 45(5):899–911.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint*. ArXiv:2009.07896 [cs].

Michalis Korakakis, Andreas Vlachos, and Adrian Weller. 2024. ALVIN: Active Learning Via INterpolation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22715–22728, Miami, Florida, USA. Association for Computational Linguistics.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2019. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, 84(5):905–949. Publisher: SAGE Publications Inc.

Punit Kumar and Atul Gupta. 2020. Active Learning Query Strategies for Classification, Regression, and Clustering: A Survey. *Journal of Computer Science and Technology*, 35(4):913–945.

Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

David D. Lewis and William A. Gale. 1994. A Sequential Algorithm for Training Text Classifiers. In *SIGIR '94*, pages 3–12, London. Springer.

Bruce G. Link and Jo Phelan. 2014. Stigma power. *Social Science & Medicine (1982)*, 103:24–32.

Bruce G. Link and Jo C. Phelan. 2001. Conceptualizing stigma. *Annual Review of Sociology*, 27:363–385. Place: US Publisher: Annual Reviews.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. ArXiv:1907.11692 [cs].

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active Learning by Acquiring Contrastive Examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Blake Miller, Fridolin Linder, and Walter 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.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Laura K. Nelson. 2021. Leveraging the alignment between machine learning and intersectionality: Using word embeddings to measure intersectional experiences of the nineteenth century U.S. South. *Poetics*, 88:101539.

Bernice A. Pescosolido, Andrew Halpern-Manners, Liying Luo, and Brea Perry. 2021. Trends in Public Stigma of Mental Illness in the US, 1996-2018. *JAMA network open*, 4(12):e2140202.

Jo C. Phelan, Bruce G. Link, and John F. Dovidio. 2008. Stigma and prejudice: One animal or two? *Social Science & Medicine*, 67(3):358–367.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Pedro L. Rodriguez, Arthur Spirling, and Brandon Stewart. 2023a. *conText: 'a la Carte' on Text (ConText) Embedding Regression*. R package version 1.4.3.

Pedro L. Rodriguez, Arthur Spirling, and Brandon M. Stewart. 2023b. Embedding Regression: Models for Context-Specific Description and Inference. *American Political Science Review*, 117(4):1255–1274.

Nicholas Roy and Andrew McCallum. 2001. Toward Optimal Active Learning through Sampling Estimation of Error Reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 441–448, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Georg Schomerus, Stephanie Schindler, Christian Sander, Eva Baumann, and Matthias C. Angermeyer. 2022. Changes in mental illness stigma over 30 years - Improvement, persistence, or deterioration? *European Psychiatry: The Journal of the Association of European Psychiatrists*, 65(1):e78.

Christopher Schröder, Lydia Müller, Andreas Niekler, and Martin Potthast. 2023. Small-Text: Active Learning for Text Classification in Python. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 84–95. ArXiv:2107.10314 [cs].

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting Uncertainty-based Query Strategies for Active Learning with Transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203, Dublin, Ireland. Association for Computational Linguistics.

M. Sittner, T. Rechenberg, S. Speerforck, M. C. Angermeyer, and G. Schomerus. 2024. 'Broken souls' vs. 'mad ax man' – changes in the portrayal of depression and schizophrenia in the German media over 10 years. *Epidemiology and Psychiatric Sciences*, 33:e37.

Dustin S. Stoltz and Marshall A. Taylor. 2021. Cultural cartography with word embeddings. *Poetics*, 88:101567.

Dustin S. Stoltz and Marshall A. Taylor. 2024. *Mapping texts: computational text analysis for the social sciences*. Computational social science. Oxford University Press, New York, NY.

Dustin S. Stoltz, Marshall A. Taylor, and Jennifer S. K. Dudley. 2024. A Tool Kit for Relation Induction in Text Analysis. *Sociological Methods & Research*, page 00491241241233242. Publisher: SAGE Publications Inc.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient Few-Shot Learning Without Prompts. *arXiv preprint*. ArXiv:2209.11055.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need.

Francisco Viveros-Jiménez, Alexander Gelbukh, and Grigori Sidorov. 2013. Simple Window Selection Strategies for the Simplified Lesk Algorithm for Word Sense Disambiguation. In *Advances in Artificial Intelligence and Its Applications*, pages 217–227, Berlin, Heidelberg. Springer.

Yue Wang, Kai Zheng, Hua Xu, and Qiaozhu Mei. 2018. Interactive medical word sense disambiguation through informed learning. *Journal of the American Medical Informatics Association*, 25(7):800–808.

Elisa M. Wirsching, Pedro L. Rodriguez, Arthur Spirling, and Brandon M. Stewart. 2025. Multilanguage Word Embeddings for Social Scientists: Estimation,

Inference, and Validation Resources for 157 Languages. *Political Analysis*, 33(2):156–163.

World Health Organization. 2019/2021. *International Classification of Diseases, Eleventh Revision (ICD-11)*. WHO.

Shujian Zhang, Chengyue Gong, Xingchao Liu, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. 2022. ALLSH: Active Learning Guided by Local Sensitivity and Hardness. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1328–1342, Seattle, United States. Association for Computational Linguistics.

Jingbo Zhu and Eduard Hovy. 2007. Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 783–790, Prague, Czech Republic. Association for Computational Linguistics.

# A Appendix

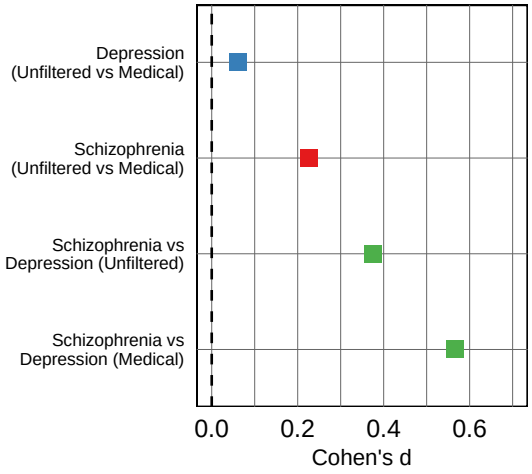


Figure A.1: Cohen's *d* effect sizes between full corpus and medical articles grouped by disease.

| Schizophrenia | | | Depression | | |
|---|---|---|---|---|---|
| $N$ | Minority F1 | Macro F1 | $N$ | Minority F1 | Macro F1 |
| 50 | 0.43 | 0.65 | 50 | 0.00 | 0.47 |
| 75 | 0.64 | 0.77 | 100 | 0.38 | 0.66 |
| 125 | 0.80 | 0.86 | 125 | 0.56 | 0.76 |
| 225 | 0.69 | 0.80 | 150 | 0.69 | 0.82 |
| 250 | 0.75 | 0.83 | 175 | 0.66 | 0.81 |
| 275 | 0.77 | 0.85 | 225 | 0.69 | 0.83 |
| 300 | 0.78 | 0.86 | 325 | 0.66 | 0.81 |
| 350 | 0.81 | 0.87 | 350 | 0.80 | 0.89 |

Table A.1: Active learning history for schizophrenia and depression

| Schizophrenia | | | Depression | | |
|---|---|---|---|---|---|
| $N$ | Minority F1 | Macro F1 | $N$ | Minority F1 | Macro F1 |
| 50 | 0.50 | 0.69 | 50 | 0.00 | 0.47 |
| 150 | 0.57 | 0.73 | 150 | 0.47 | 0.71 |
| 250 | 0.57 | 0.73 | 250 | 0.53 | 0.74 |
| 350 | 0.67 | 0.79 | 350 | 0.67 | 0.81 |

Table A.2: Random sampling learning history for schizophrenia and depression

| Schizophrenia | | | Depression | | |
|---|---|---|---|---|---|
| $N$ | Minority F1 | Macro F1 | $N$ | Minority F1 | Macro F1 |
| 50 | 0.42 | 0.64 | 50 | 0.00 | 0.47 |
| 150 | 0.28 | 0.48 | 150 | 0.14 | 0.50 |
| 250 | 0.39 | 0.52 | 250 | 0.12 | 0.42 |
| 350 | 0.34 | 0.49 | 350 | 0.23 | 0.45 |

Table A.3: Cosine similarity bootstrapping learning history for schizophrenia and depression

| Disease | F1 | Batch Size | Learning Rate | Max Steps | Num Iterations |
|---|---|---|---|---|---|
| Depression | 0.833 | 16 | 8.39e-06 | 260 | 30 |
| Depression | 0.833 | 16 | 6.87e-06 | 260 | 30 |
| Schizophrenia | 0.830 | 32 | 1.16e-05 | 260 | 50 |
| Schizophrenia | 0.830 | 16 | 2.71e-05 | 80 | 20 |

Table A.4: Training parameter configurations for the best performing models for schizophrenia and depression.

| Comparison | Cohen's $d$ [95% CI] | Interpretation |
|---|---|---|
| Depression (Unfiltered vs Medical) | 0.06 [0.057, 0.066] | Small Effect |
| Schizophrenia (Unfiltered vs Medical) | 0.23 [0.216, 0.237] | Small to Medium Effect |
| Schizophrenia vs Depression (Unfiltered) | 0.376 [0.368, 0.383] | Medium Effect |
| Schizophrenia vs Depression (Medical) | 0.565 [0.556, 0.573] | Medium to Large Effect |

Table A.5: Cohen's $d$ comparing the full corpus and medical documents between groups.

| Disease | Predicted Class | n | Percent | Label |
|---|---|---|---|---|
| Schizophrenia | 0 | 24,945 | 26.1 | Non-Medical |
| Schizophrenia | 1 | 70,784 | 73.9 | Medical |
| Depression | 0 | 56,407 | 13.2 | Non-Medical |
| Depression | 1 | 370,638 | 86.8 | Medical |

Table A.6: Label distributions of the final classifiers trained with active learning on the full corpus for each schizophrenia and depression.

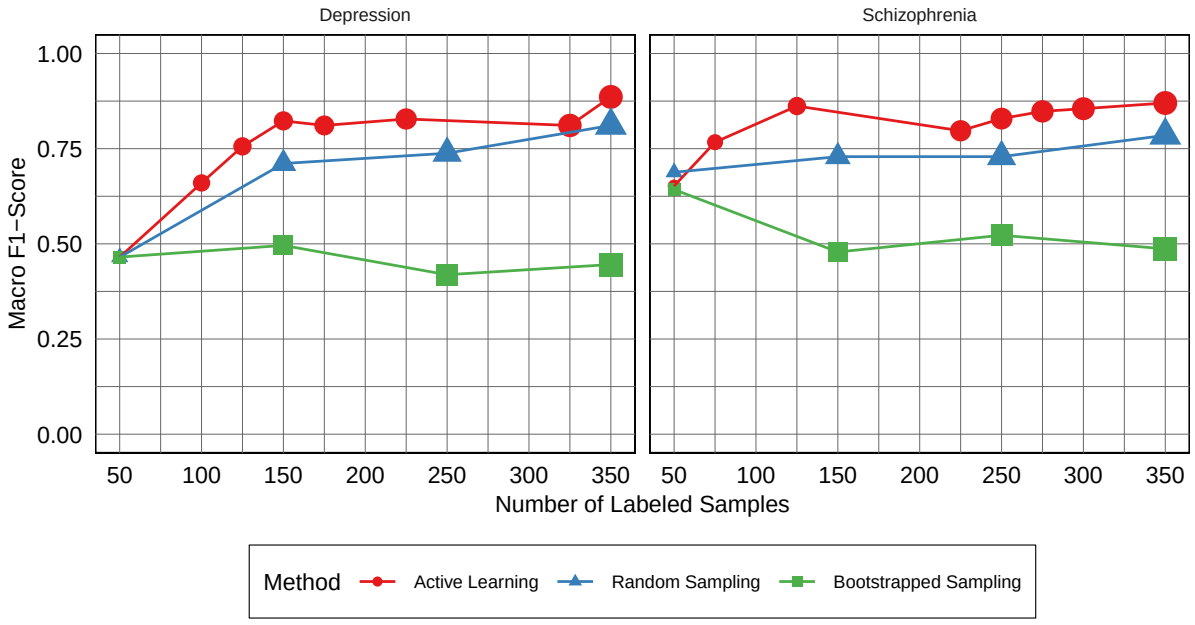


Figure A.2: Evolution of the Macro F1-score on the test-set by number of labeled examples and sampling method for schizophrenia and depression.