Original Paper

Semantic Similarity Search Approach to Extract Exemplars of Stigmatizing and Positive Language in Obstetric Clinical Notes: Exploratory Study

Jihye Kim Scroggins[1], PhD, RN; Ismael Ibrahim Hulchafo[2], MD, MS; Veronica Barcelona[2], PhD, RN; Anahita Davoudi[3], PhD; Hans Moen[4], PhD; Sarah Harkins[2], PhD, RN; Danielle Scharp[5], PhD, APRN, FNP-BC; Kenrick Cato[6], PhD, RN, CPHIMS, FAAN; Michele Tadiello[7], MMCi, BSN; Maxim Topaz[2], PhD, RN

[1]School of Nursing, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[2]School of Nursing, Columbia University, New York, NY, United States
[3]VNS Health, New York, NY, United States
[4]Department of Computer Science, Aalto University, Espoo, Finland
[5]Icahn School of Medicine at Mount Sinai, New York, NY, United States
[6]School of Nursing, University of Pennsylvania, Philadelphia, PA, United States
[7]Center for Community-Engaged Health Informatics and Data Science, Columbia University Irving Medical Center, New York, NY, United States

Corresponding Author:

Jihye Kim Scroggins, PhD, RN
School of Nursing, University of North Carolina at Chapel Hill
120 Medical Drive
Chapel Hill, NC 27514
United States
Phone: 1 (919) 966-4260
Email: jihyeks@unc.edu

Abstract

Background: Natural language processing can extract meaningful information from clinical notes. However, human annotation is time-consuming and costly, and scarce data poses a challenge.

Objective: This study aimed to explore a semantic similarity search approach to extract exemplars of stigmatizing and positive language in obstetric clinical notes.

Methods: We used electronic health record data from labor and birth admissions at 2 hospitals in the United States from 2017 to 2019. We used a semantic similarity search approach, which used 200 randomly selected true exemplars, stratified by language categories, as queries to search for similar exemplar candidates. We extracted the top 5 candidates with the highest cosine similarities, which were assessed for accuracy.

Results: We retrieved 1000 candidates. An average precision of 0.69 was achieved when candidates with cosine similarity thresholds of 0.75 or higher were included, at which point 68.8% (64/93) of exemplar candidates accurately represented true cases. At the 0.75 threshold, the proportion of true cases was higher for preferred language (41/56, 73.2%) and unilateral/authoritarian decisions (5/7, 71.4%). The proportion of true cases was lower for difficult patients (2/7, 28.6%) and marginalized identities (3/9, 33.3%).

Conclusions: The semantic similarity search approach shows promise in efficiently extracting exemplars while reducing the annotation burden, laying the groundwork for future applications in other domains.

JMIR Med Inform 2026;14:e85088; doi: 10.2196/85088

Keywords: natural language processing; electronic health records; stigmatizing language; health communication; bias

Introduction

Natural language processing (NLP) can efficiently and accurately extract meaningful information from large text data, such as clinical notes within electronic health records (EHRs) [1,2]. One significant challenge is obtaining sufficient human-annotated data to train NLP models for specific tasks and datasets. Human annotation is labor-intensive and

costly, requiring extensive effort to review and label data for exemplars representing specific concepts or categories within the text. For example, annotating 50 discharge summaries can take approximately 80 hours for a team of several clinical experts; this time grows exponentially for datasets with a low prevalence of the target information [3]. Given the substantial time and manpower needed for human annotation, it is common to encounter limited human-labeled data for NLP model development [4].

Prior NLP research has focused on reducing annotation burden through approaches such as active learning and data augmentation. Traditional active learning, most commonly using uncertainty-based querying, supports annotation by iteratively selecting unlabeled cases for which the model has the highest uncertainty (lowest confidence) and prioritizing these cases for human review [5]. However, uncertainty-based querying may be less effective for low-prevalence or rare concepts, which can be under-selected when models exhibit high confidence in assigning majority-class labels; for example, rare positive cases may be confidently misclassified as negative and therefore not selected for human review [6,7]. Data augmentation, including synonym replacement, paraphrasing, and more recently, synthetic data generation using large language models, has also been used to expand training datasets [8,9]. However, these approaches may unintentionally introduce noise or alter subtle linguistic or contextual cues essential for capturing nuanced context-dependent concepts [10,11]. For example, human evaluation of synthetic data quality in the context of stigmatizing language has shown that synthetic exemplars exhibited limited clinical realism [12].

Semantic similarity search offers a complementary strategy for expanding training data by efficiently extracting additional exemplars that are semantically similar to existing human-annotated exemplars. Semantic similarity refers to the degree to which 2 pieces of text share similar meanings based on their linguistic content and context [13]. Semantic similarity has been widely studied. Early approaches focused on document-level similarity based on lexical and distributional overlap [14]. More recent work has focused on sentence-level similarity using neural embeddings, which enable more nuanced comparisons across short text spans [15]. In the clinical domain, prior work has largely focused on developing and evaluating models for estimating similarity between pairs of clinical sentences [16,17]. These studies typically use trained models to predict similarity scores between sentence pairs and evaluate how closely these predictions align with human ratings [16,17]. Semantic similarity has also been used in information retrieval and retrieval-augmented generation systems to retrieve clinically relevant texts, passages, or documents [18,19]. These approaches are primarily used for downstream tasks such as question answering and clinical decision support [20]. In contrast, the use of semantic similarity search to support annotation workflows, specifically for retrieving additional exemplars, remains relatively limited.

In the current study, we applied semantic similarity search to support annotation by using human-annotated true exemplars as queries to retrieve additional exemplar candidates. This approach has the potential to reduce the human labor associated with reviewing and labeling new text data from scratch. Unlike traditional active learning, this approach directly retrieves semantically similar exemplar candidates to human-annotated true exemplars, without relying on iterative model training or uncertainty-based querying. This strategy may be particularly useful for identifying rare concepts, as it can retrieve semantically similar text even when such cases are infrequent, allowing expansion of sparse categories. In addition, by using true exemplars as queries, this approach can preserve real-world language and subtle contextual features that may not be well captured by synthetic data.

Specifically, we explored the application of this approach in the context of identifying stigmatizing and positive language categories in obstetric clinical notes. Stigmatizing language can convey implicit or explicit bias [21], which can negatively influence patient-clinician relationships and satisfaction in health care [22,23]. Identifying and mitigating stigmatizing language is essential for providing respectful and unbiased care for perinatal populations affected by significant health disparities [24]. The prevalence of stigmatizing language can be as low as 1% to 2% for certain language categories or patient populations [25,26]. This presents challenges for manual annotation and obtaining sufficient data for optimal NLP model training. Additionally, identifying positive language that respects patients' views and autonomy is important to promote the use of strength-based language in clinical documentation [27]. Our efforts to identify stigmatizing and positive language revealed the need for an efficient approach to extract additional exemplars with limited resources [28]. This study aimed to evaluate semantic similarity search as a more efficient approach for identifying additional exemplars of stigmatizing and positive language in obstetric clinical notes. We assessed retrieval accuracy through human evaluation and precision metrics.

Methods

Study Data

We used EHR data from patients at >20 weeks' gestation admitted for labor and birth at 2 urban hospitals in the Northeast United States from 2017 to 2019. All clinical notes in the EHR during the inpatient hospital stay were eligible. Note types with limited clinician narrative text describing patient assessments or impressions, such as medication orders, transfer notes, and template-based statements about procedures or operations, were excluded. Seven note types were used in the current study: obstetric postpartum note, obstetric admission note, obstetric triage note, anesthesia resident note, miscellaneous nursing note, social work initial assessment, and initial nutrition assessment.

Ethical Considerations

We received institutional review board (IRB) approval from Columbia University Medical Center (AAAT9870) for this study under expedited review, category 5: research involving

materials (data, documents, records, or specimens) that have been collected or will be collected solely for nonresearch purposes, such as medical treatment or diagnosis. A waiver of informed consent was granted by the IRB because the research involved no more than minimal risk, did not adversely affect the rights and welfare of human subjects, could not be carried out without the waiver, and the study could not be completed without using identifiable private information from patients discharged from the study hospitals. To protect privacy and confidentiality, study data were stored on a secure server following institutional data security guidelines, with access restricted to authorized study personnel with a direct need for data management and analysis. All study personnel completed IRB-required training in the responsible conduct of research and protection of human subjects. No compensation was provided because this study involved secondary use of data without direct participant involvement, interaction, or recruitment.

Semantic Similarity Search Approach

Approach Overview

We explored using semantic similarity search to expand the initial human-annotated dataset in a less manually intensive way. This approach uses initial human-annotated exemplars (ie, *true exemplars*) as queries to search for similar exemplars (ie, *exemplar candidates*). Then, human experts can perform a relatively faster review of the exemplar candidates to determine the accuracy.

Language Categories

Stigmatizing and positive language categories are presented in Table 1. The categories and operational definitions were developed through iterative qualitative analysis [26,29] and informed by prior work on stigmatizing language [30]. A multidisciplinary team of experts, including obstetrics and gynecology physicians, nurses, and nonclinical researchers, contributed diverse perspectives to the development of these categories. These efforts were intended to enhance the rigor of the operational definitions.

Table 1. Stigmatizing and positive language categories.

Language category	Definition	Examples	True exemplars (N=200), n
Stigmatizing language			
Difficult patient	• Nonadherence or noncompliance with plan of care, refusal of care, referrals, or services.	• "missed nutrition appt again says she doesn't want to go." • "exam unchanged however now c/o [complains of] ctx [contraction] pain." • "poor compliance with antenatal visits."	28
Unilateral/ authoritarian decisions	• Language that supports clinician's authority over patients. Upholds hierarchy centering clinician, not patient.	• "cervix unfavorable therefore will start induction with Cytotec." • "SW [social worker] has advised pt [patient] that if there continues to be yelling in room, ACS [adult and child services] will need to be contacted." • "also very upset about persistent shaking in her UE [upper extremities]. Explained to patient multiple times that shaking is normal and a result of hormones and the anesthesia."	14
Power/privilege	• Describes power and privilege identities related to psychological or social-ecological status.	• "pt [patient] reports having a nurturing marriage with FOB [father of the baby] who works as a Lawyer." • "private pt [patient] of mine at 39+6 wks [weeks] with multiple episodes of emesis this AM." • "Caucasian female. pleasantly engaged in conversation."	17
Marginalized identities	• Documentation of social and behavioral risk factors in narrative that could contribute to marginalization. • Restating for emphasis that is already in the checklist/form data or unnecessary patient descriptor (eg,	• "social disarray, estranged from family with FOB [father of the baby] not involved." • "Dominican female. Initial psychosocial assessment due to teen pregnancy late registrant."	74

Language category	Definition	Examples	True exemplars (N=200), n
		"toxic habits," "financially supports self," "teen mother," "late registrant," and "obesity").	• "Patient is a 32yo [year old] Dominican unmarried unemployed female."	
Questioning patient credibility	• Disbelief in patient's report of health and social history or status.	• "unsure if patient telling the truth." • "pt [patient] denied any other DV [domestic violence] incidents and was adamant that relationship with spouse was healthy." • "Pt [patient] does not think she has gestational diabetes she states she ate very sweet rice (a dish from her country) the day before the glucose challenge which is what she believes this is the reason for the high value states it has never been high before."	11
Disapproval	• Behaviors of patient not in alignment with health care provider expectations.	• "FOB [father of baby] is 23 yo [year old] unemployed and not as involved as he should be." • "Pt [patient] states she tried to obtain f/u [follow-up] appt [appointment]. but did not get a call back. Advised she should have gone to the clinic to straighten things out- in the future to take initiative-importance stressed as well as appts [appointments] for her newborn." • "PP [postpartum] BCM [bridge contraceptive method]- pt states she prefers to use condoms will continue to readdress."	6
Structural/interprofessional hierarchy	• Notes stating issues in care are attributed to another clinician, often nurses. Notes stating that care was not delivered in a timely manner due to staffing or other structural issues.	• "Nurse states IV [intravenous] 'infiltrated,' but it was fine." • "Unable to staff nursing for placement of epidural in triage. Requested that charge nurse please inform anesthesiology team about when there is enough nursing staff." • "Patient continues to wait in triage until L D [Labor & Delivery] bed available."	2
Positive language			
Preferred language	• Preferred words that can convey patient's point of view respectively and objectively (eg, endorses, reports, or states).	• "patient desires to ambulate and encourage natural labor." • "patient states she feels some pain on right side." • "states she has had irregular ctx [contractions] starting 24 hours ago. endorses FM [fetal movement]. Endorses N/V [nausea/vomiting]"	44
Autonomy for birth	• Notes that indicate patient exercising autonomy around birth and plan of care.	• "Pt [patient] declines epidural at this time- trying to deliver without analgesia." • "will give patient the option to have continuous monitoring overnight and give her the opportunity to go into spontaneous labor with plans for c-section tomorrow if she does not progress." • "low risk screening declined diagnostic. She was given the option of admission with possible augmentation of labor vs discharge home to labor. Plan was made for admission."	4

Initial Human Annotation

To generate the initial human-annotated dataset, 4 researchers with expertise in qualitative research and/or nursing independently annotated clinical notes following an established codebook (see Table S4 in Multimedia Appendix 1). The codebook was developed through iterative inductive-deductive content analysis [26]. All annotators were trained prior to participating in annotations, which included review of the codebook, discussion of example phrases and sentences, and coding demonstrations. Annotators manually labeled true exemplars from a total of 1771 clinical notes, which typically spanned 1 to 3 sentences. To ensure reliability, 2 annotators reviewed and annotated the same clinical notes. We resolved disagreements through iterative discussions among the annotators. The initial agreement among the annotators across the 1771 clinical notes was fair (Cohen κ=0.4, agreement rate=72%), which reflects the inherently nuanced nature of language use and subjectivity involved with interpreting complex and nuanced language [31]. We spent extensive effort and time discussing any discrepancies among annotators to reach a consensus and ensure the quality and accuracy of the final annotated dataset [31].

Sentence-Transformer Model

We used a sentence-transformer model, "multi-qa-distilbert-cos-v1," from the Hugging Face Model Hub [15,32] to search for additional exemplar candidates that were semantically similar to the set of true exemplars. This model is specifically designed for semantic similarity search and retrieval tasks using a contrastive learning objective to generate sentence-level embeddings that enable comparison of a given sentence with other sentences that are most closely related in meaning [32]. This design enables efficient, out-of-the-box application of the model for similarity-based retrieval tasks, making it well suited for the current task. Although this model is not trained on clinical text, it is trained on large-scale, question-answer-style data from diverse sources [32], which contributes to capturing contextual meaning across varied linguistic expressions. This general-domain training may be helpful for identifying broader sociolinguistic patterns and contextual features in clinical text that can extend beyond clinical language. Prior work suggests that general-domain embedding models can perform well for semantic similarity tasks in clinical text, although performance may vary across specific models and tasks [33].

Exemplar candidates were selected from clinical notes not used in the initial human annotation. We preprocessed the clinical notes by converting the free-text portions into a format similar to the true exemplars (single, double, and triple consecutive sentences). Sentence tokenization was performed first using natural language toolkit's "sent_tokenize," which identifies sentence boundaries based on learned punctuation and linguistic patterns rather than simple rule-based splitting. After tokenization, sliding windows of 1-, 2-, and 3-sentence spans were generated by advancing one sentence at a time. Duplicate candidates were then detected using a 6-word prefix and removed in 2 stages: first keeping only the earliest instance of each prefix, and then alternately dropping any remaining repeats to ensure only 1 representative remained. No normalization of casing, numbers, or abbreviations was applied. All text was embedded in its original form to preserve real-world clinical text. Examples of synthetic, free-text note snippets can be found in Table S3 in Multimedia Appendix 1.

We used the sentence-transformer model to encode true exemplars into fixed-size vectorized embeddings [15]. Then, true exemplars were used as queries to search and recommend new exemplars that were likely candidates from the unused clinical note datasets. We calculated semantic similarity between the true exemplars and exemplar candidates using the cosine similarity applied to their associated embedding representations. The Facebook AI similarity search library was used to perform the similarity searches [34]. We stratified and randomly selected 200 true exemplars across language categories to be used as queries. Stratification was conducted by language category to preserve each category's natural prevalence in the initial human-annotated dataset. As a result, less frequent categories contributed fewer exemplars (Table 1). For each of the 200 true exemplars, the top 5 similar exemplar candidates with the highest cosine similarities were retrieved (N=1000 exemplar candidates) for human review. Pseudocode blocks for the semantic similarity search pipeline employed in this study are provided in Multimedia Appendix 2.

Examining Accuracy of Exemplar Candidates

Four researchers who performed the initial annotation independently reviewed and assessed the exemplar candidates. Reviewers labeled the exemplar candidate as a "match" if it accurately represented the corresponding language category and as "not a match" if it did not. We resolved disagreements through iterative discussions among reviewers to reach consensus. We calculated the frequency and percentage of matched and nonmatched cases as true positives and false positives. We examined how precision changes when different cosine similarity boundaries were used to inform optimizing this approach for future research and applications. We calculated and reported average precision and corresponding 95% Wilson CIs across different cosine similarity thresholds, ranging from 0.5 to 0.95 in increments of 0.05. To our knowledge, there is no universally accepted precision cutoff for information retrieval tasks to support annotation workflows. Information retrieval literature supports that performance should be evaluated in relation to the intended use and context of each task, as performance requirements may vary across applications [35]. Previous research on annotation support often focuses on reducing annotation burden while maintaining annotation quality [36,37]. In a recent study on automated annotation support, the authors noted that precision exceeding 0.7 was considered "strong performance" for their task [38]. Given the exploratory nature of this study and its intended use to support annotation workflows, we selected precision ≥0.70 as an acceptable, pragmatic point to balance accuracy with meaningful efficiency gains in reducing annotation burden.

At this precision point, most exemplar candidates would represent true cases while allowing a manageable proportion of false positives.

Results

A total of 1000 exemplar candidates were retrieved using the semantic similarity search approach (Figure 1). These exemplar candidates were derived from 443 clinical notes of 403 patients. The number of exemplar candidates by note type is provided in Table 2. Among the 1000 exemplar candidates retrieved, most were derived from obstetric admission notes (n=295, 29.5%), followed by obstetric postpartum notes (n=202, 20.2%) and miscellaneous nursing notes (n=172, 17.2%). Human annotators independently reviewed all 1000 exemplar candidates retrieved for accuracy. The interrater reliability among the annotators was good (Cohen $\varkappa$=0.71, 95% CI 0.67-0.75) across 1000 exemplar candidates.

Figure 1. Flow diagram summarizing the number of true exemplars or exemplar candidates in the analysis and results.

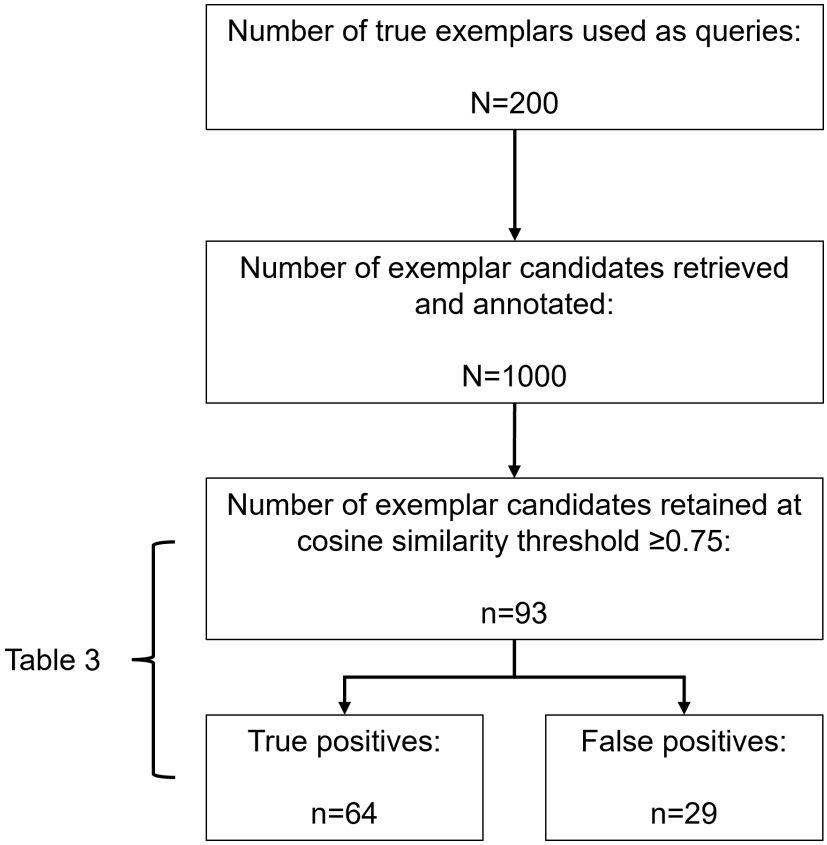


Table 2. Number of exemplar candidates by note type.

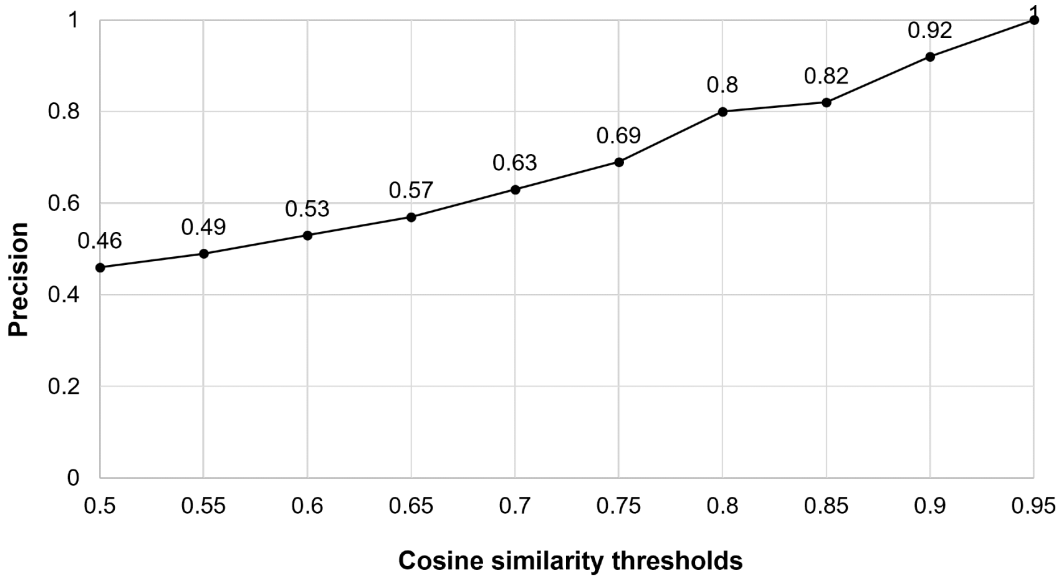
Note type	Exemplar candidates[a], n (%)
Obstetric admission note	295 (29.5)
Obstetric postpartum note	202 (20.2)
Miscellaneous nursing note	172 (17.2)
Anesthesia resident note	137 (13.7)
Obstetric triage note	106 (10.6)
Social work initial assessment	51 (5.1)
Initial nutrition assessment	37 (3.7)
Total exemplar candidates retrieved	1000 (100)

[a]The total number of exemplar candidates retrieved was used as the denominator to calculate the percentages.

We examined how the average precision of all language categories changed when including exemplar candidates at different cosine similarity thresholds. As shown in Figure 2, average precision increased with higher cosine similarity thresholds. The number of exemplar candidates retained decreased with higher cosine similarity thresholds. Average precision was highest at 1 (95% CI 0.44-1.00) when including exemplar candidates with cosine similarity ≥0.95, at which point 3 of 3 exemplar candidates were true cases. Average precision was lowest at 0.46 (95% CI 0.43-0.49) when including exemplar candidates with cosine similarity ≥0.5, at which point 383 of 831 exemplar candidates were true

cases. Acceptable average precision (≥0.70) was achieved when including exemplar candidates with a cosine similarity of at least 0.75.

Figure 2. Line graph illustrating the average precision at different cosine similarity thresholds with corresponding values for the total number of exemplar candidates retained at each threshold, the number of true positives, and 95% CIs. The total number of exemplar candidates retained at each threshold was used as the denominator to calculate the precision.



	Cosine similarity threshold									
	0.5	**0.55**	**0.6**	**0.65**	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**
Exemplar candidates retained at threshold	831	677	466	302	201	93	44	28	13	3
True positives (Matched cases)	383	333	245	173	127	64	35	23	12	3
95% CIs	0.43-0.49	0.45-0.53	0.48-0.57	0.52-0.63	0.56-0.70	0.59-0.77	0.65-0.89	0.64-0.92	0.67-0.99	0.44-1.00

A total of 93 exemplar candidates had a cosine similarity of at least 0.75. Among those, 64 (68.8%) accurately represented the true cases across stigmatizing and positive language categories (Table 3). In specific language categories, 100% (5/5) of exemplar candidates for autonomy for birth, 73.2% (41/56) for preferred language, and 71.4% (5/7) for unilateral/authoritarian decisions represented the true cases. While the proportions of matched cases were high for power/privilege, questioning patient credibility, and disapproval, the number of retrieved exemplar candidates for these categories was small at the 0.75 cosine similarity threshold (eg, 1 of 1 exemplar candidate was a true case for power/privilege and disapproval, respectively). In contrast, 28.6% (2/7) of exemplar candidates for difficult patient and 33.3% (3/9) of exemplar candidates for marginalized identities represented the true cases, with lower proportions of matched cases than other language categories at the 0.75 cosine similarity threshold.

Table 3. Proportion of matched cases by language category at a cosine similarity threshold of 0.75.

Language category	Exemplar candidates, N[a]	True positives (matched cases), n (%)	False positives (nonmatched cases), n (%)
Preferred language	56	41 (73.2)	15 (26.8)
Autonomy for birth	5	5 (100)	0 (0)
Difficult patient	7	2 (28.6)	5 (71.4)
Unilateral/authoritarian decisions	7	5 (71.4)	2 (28.6)
Power/privilege	1	1 (100)	0 (0)
Structural/interprofessional hierarchy	4	3 (75)	1 (25)
Marginalized identities	9	3 (33.3)	6 (66.7)
Questioning patient credibility	3	3 (100)	0 (0)

Language category	Exemplar candidates, N[a]	True positives (matched cases), n (%)	False positives (nonmatched cases), n (%)
Disapproval	1	1 (100)	0 (0)
Total	93	64 (68.8)	29 (31.2)

[a]The total N column was used as the denominator to calculate the proportion of matched and nonmatched cases.

Detailed results for precision at different cosine similarity thresholds by specific language categories and corresponding CIs are reported in Tables S1 and S2 in Multimedia Appendix 1. Of note, several precision values were based on small numbers of exemplar candidates, particularly at higher thresholds, and are statistically unstable for these sparse categories. For categories with relatively larger numbers of retained candidates, such as preferred language, acceptable precision (≥0.70) was observed at a cosine similarity of 0.70, where 74 of 105 exemplar candidates represented true cases. False positives most commonly occurred in exemplars containing more complex social or contextual information. For example, exemplars describing objective social circumstances were misclassified as stigmatizing despite the absence of judgmental language (eg, false positive in the marginalized identities category: "Patient may want to be discharged to home tonight. She lives nearby and will be able to visit infant"). In other cases, exemplars containing stigmatizing language were misclassified into incorrect categories. For example, "Patient stated she doesn't have any help in the country and FOB [father of baby] is in DR [Dominican Republic]. Patient needs social worker, will continue to monitor..." was incorrectly classified into the power/privilege category.

Discussion

Principal Findings

We explored a semantic similarity search approach to identify additional exemplars from obstetric clinical notes with less manual human effort. We found that we can increase accuracy by adjusting cosine similarity boundaries when selecting exemplar candidates. The average precision increased with higher cosine similarity thresholds, achieving an acceptable average precision ≥0.70 when exemplar candidates with a cosine similarity of 0.75 or higher were included. On average, 68.8% (64/93) of exemplar candidates accurately represented true cases across various language categories at the 0.75 cosine similarity threshold. These thresholds should be guided by each use case in future research and application. Lower thresholds may be useful for exploratory analyses, where broader retrieval is desired to examine patterns of expressions and language use. Lower thresholds may also be appropriate when maximum coverage is prioritized over precision, such as for rarely occurring cases, where missing potential matches would be more consequential than reviewing additional false positives. In contrast, higher thresholds may be suitable when precision should be prioritized, such as using retrieved texts for direct application in clinical settings, including quality review or clinical documentation audit to identify potential stigmatizing language. Higher thresholds may also be useful with large datasets, where even a strict precision cutoff can yield a sufficient number of exemplar candidates, whereas the same threshold applied to smaller datasets may result in too few retrieved exemplars to be useful.

It is important to note that the reported accuracy in our study reflects only positive retrieval quality, measured by precision, rather than the overall effectiveness or completeness of the retrieval approach. Given the exploratory nature of the study, we did not annotate nonretrieved texts to calculate more comprehensive evaluation metrics, including recall and F_1-scores. As such, our findings do not provide information regarding the extent to which the approach identified all relevant exemplars within the dataset. For example, it remains unknown how many relevant exemplars may not have been retrieved, limiting the ability to assess the completeness of retrieval. In addition, the number of exemplar candidates retained decreased substantially across language categories as cosine similarity thresholds increased (detailed category-specific results in Multimedia Appendix 1). As such, the current study may be underpowered to support reliable category-level conclusions. When few exemplar candidates were retained, high precision values should not be misinterpreted as strong performance, as they are statistically unstable and reflect limited sample size. At lower cosine thresholds, where more candidates were retained, precision varied across language categories. Although these should be interpreted cautiously, they may reflect different language complexity. For example, at the 0.5 cosine similarity threshold, higher precision was observed for preferred language and autonomy for birth compared to other categories. These 2 categories were more straightforward with well-defined, identifiable keywords (eg, "desires natural birth," "reports pain"). Conversely, we observed lower precision for marginalized identities, questioning patient credibility, and disapproval categories at the 0.5 cosine similarity threshold. These categories were more nuanced and context-dependent. For example, there was a subtle undertone of disapproving the patient's choice of birth control in the following: "postpartum birth control method-patient states that she prefers to use condoms and will continue to readdress."

Accurately identifying such nuances requires a deeper understanding and interpretation of contextual meanings and subtleties, which may be challenging for semantic similarity search. Semantic similarity search primarily relies on vector-based representations of text, including static embeddings like Word2Vec and GloVe [13,39] and contextual embeddings like bidirectional encoder representations from transformers (BERT) [40]. Static embeddings can

effectively capture general semantic relationships [13,39] but can be limited in detecting context-specific nuances and meanings [41]. Although contextual embeddings, such as those generated by the sentence-transformer model used in the current study, can better capture contextual information, they may still be challenged by complex and subtle nuances [42]. Finally, when suitable training data are available, sentence-transformer models could be further improved through data- and task-specific training by fine-tuning.

While semantic similarity has been previously used for tasks such as information retrieval, its application to support annotation workflows, specifically for retrieving additional exemplars, remains relatively limited. The current study extends this line of work by applying semantic similarity search to more efficiently expand annotated datasets. This approach is particularly useful for low-prevalence concepts, where traditional active learning strategies may be less effective. In contrast to data augmentation strategies, such as synthetic data generation, it can preserve naturally occurring language while enabling targeted expansion of context-dependent concepts. Furthermore, while prior retrieval-based approaches have used similarity to identify relevant text for tasks such as information retrieval or retrieval-augmented generation, the effectiveness of these approaches is typically evaluated based on automated metrics or downstream model performance, rather than human validation [18,19]. In contrast, we conducted human evaluation of retrieved exemplar candidates and examined precision across cosine similarity thresholds, providing insight into the utility of this approach as an annotation support strategy. Our findings highlight how cosine similarity thresholds influence precision, offering practical guidance for applying this approach in settings where annotated data are sparse.

Although our findings are based on the context of stigmatizing and positive language in obstetric clinical notes, which may limit the generalizability to other specialized datasets or domains, semantic similarity search itself is not inherently domain-specific. Because this approach operates by comparing a given sentence with other sentences to identify those most closely related in meaning, it is not dependent on or limited to specific models but can be incorporated into existing clinical NLP pipelines as a data augmentation or annotation-support step. For example, retrieved exemplar candidates may be used to efficiently expand training datasets prior to NLP model development or to support targeted human review in resource-constrained settings. Prior research has demonstrated the practical benefits of targeted text selection using active learning to support annotation workflows. For example, active learning approaches have been shown to reduce annotation time by approximately 20% to 35% for clinically explicit concepts, such as medical problems and tests, by prioritizing informative texts for manual annotation [43,44]. More recently, active learning has also been shown to reduce other forms of annotation burden. For example, Nachtegael et al [45] used active learning to select unlabeled texts for manual annotation. They found that models trained on selectively annotated datasets achieved performance comparable to models trained on fully labeled datasets while reducing the amount of labeled data required by 6% to 38% [45]. Although these studies employed different active learning approaches than the current study, they indicate that strategically selecting texts for manual review can reduce annotation burden.

Limitations

This study has some limitations. First, the effectiveness of semantic similarity search depends on the quality of the initial human-annotated dataset. If true exemplars do not accurately reflect positive cases, the extracted candidates are also likely to be inaccurate. The relatively modest initial agreement among annotators (Cohen κ=0.4) highlights the subjective nature of identifying stigmatizing language. Consequently, some retrieved exemplars may have been classified differently among annotators, which could have influenced downstream retrieval accuracy. Additionally, if true exemplars do not represent diverse language categories, the extracted candidates may not be comprehensive. Second, while semantic similarity search reduced the labor required for annotation, it does not eliminate the need for human review. Due to the risk of false positives, a rapid human review of the retrieved candidates remains essential to validate the accuracy. Third, our evaluation was limited to precision to prioritize annotation efficiency. In our approach, the model retrieved exemplars predicted to be positive matches to the true exemplars. We did not annotate nonretrieved texts to identify false negatives, which is required to calculate recall. As such, the expanded dataset should be useful for augmenting existing annotated datasets with positive cases, rather than providing exhaustive or population-representative datasets. Fourth, although language categories and operational definitions were developed through qualitative analysis and informed by prior research, several categories remain inherently subjective and context-dependent. The operational definitions for these categories may remain somewhat ambiguous in practice, which can introduce variability in interpretation even among trained annotators. This reflects broader challenges in objectively operationalizing stigmatizing language in clinical notes, and findings should be considered given this inherent measurement limitation. While all discrepancies were resolved through consensus to establish the final annotated dataset, some degree of uncertainty may remain. As such, the retrieved exemplars may reflect alignment with the study-specific definitions and exemplars rather than universally accepted definitions. Fifth, we used a model trained on general-domain data because it is designed for similarity-based retrieval tasks [32], enabling its direct and efficient application in the current study without additional modifications. In addition, stigmatizing and positive language often requires capturing broader sociolinguistic expressions and context beyond clinical language, for which training on diverse data sources may be helpful. However, clinical language often includes specialized terminology, abbreviations, and domain-specific cues, which may not be fully captured by general-domain embedding models. As such, semantic similarity estimates may be less accurate for certain clinical expressions. Domain-specific models (eg, ClinicalBERT or BioClinicalBERT) are

trained on clinical text and may provide better representations of clinical terminology and context [46]. However, these models are not specifically designed or explicitly optimized for sentence-level similarity search or retrieval tasks. Thus, additional methodological adaptations, such as deriving sentence embeddings and optimizing the models for similarity-based retrieval tasks, would be required to apply domain-specific models to the task explored in the current study. For example, ClinicalBERT produces token-level contextual representations rather than sentence-level embeddings that can be directly compared using cosine similarity to identify semantically similar text. Token-level representations need to be transformed into sentence-level representations using a pooling strategy (eg, mean or max pooling), and different strategies may yield varying retrieval accuracy. Consequently, a meaningful empirical comparison with domain-specific models would be challenging because differences in retrieval accuracy could reflect these adaptations rather than the underlying models themselves. Therefore, rigorous comparison of clinical embedding models was considered beyond the scope of the current exploratory study. Finally, several language categories had very small sample sizes at higher cosine similarity thresholds, resulting in high but statistically unstable precision that should not be misinterpreted as strong or reliable performance.

Future Research

Future research may consider evaluating recall and F_1-scores, which can provide additional insight into semantic similarity search performance, particularly given the potential trade-off between precision and recall [47]. Exploring additional techniques to refine the candidate selection process may further contribute to optimizing accuracy. For instance, combining other similarity metrics, such as pragmatic similarity, could provide a more comprehensive assessment. Pragmatic similarity considers the intent and attitude behind the text, which could help to ensure that the intended meanings between texts are aligned [48]. In addition, comparing the performance of general-domain and clinical-domain models in similarity-based retrieval tasks for annotation support could provide further insight into how domain-specific models affect retrieval accuracy and annotation efficiency. Future work is also needed to further clarify, refine, and standardize the operationalization of these language categories to improve annotation consistency and reproducibility. Finally, application of this approach across diverse clinical settings and datasets will further inform generalizability.

Conclusions

This study applies semantic similarity search in the context of stigmatizing language to support annotation workflows. Unlike previous research using uncertainty-based querying or synthetic data augmentation, we used human-annotated true exemplars as queries to directly retrieve additional exemplars from real-world clinical notes to support targeted expansion of context-dependent concepts. Findings demonstrate the potential of semantic similarity search to efficiently extract additional exemplars and increase the volume of training data while reducing annotation burden. By optimizing the cosine similarity thresholds, we may further achieve higher precision. These findings provide a foundation for further refinement and application of this approach in other domains requiring efficient ways to extract additional exemplars to augment NLP training data.

Acknowledgments

We thank Arielle Hazi for her assistance during the revision process, including providing additional information on previously generated data and results. During the revision process, the first author used ChatGPT (GPT-5.3; OpenAI) to assist with proofreading tasks, specifically for identifying and correcting grammatical errors, under full human supervision. All outputs were reviewed and edited by the authors as needed, who take full responsibility for the content of the publication.

Funding

Columbia University Data Science Institute Seed Funds and the Gordon and Betty Moore Foundation grant (GBMF9048) supported this project.

Data Availability

Clinical data used in this study are restricted by the institutional review board and cannot be publicly shared.

Authors' Contributions

Analysis: JKS, IIH, VB, AD, SH, DS
Conceptualization: M Topaz
Data curation: IIH, KC, M Tadiello
Funding acquisition: VB, KC, M Topaz
Methodology: AD, HM, M Topaz
Supervision: VB, M Topaz
Visualization: JKS
Writing – original draft: JKS, IIH
Writing – review & editing: JKS, IIH, VB, AD, HM, SH, DS, KC, M Tadiello, M Topaz

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary tables presenting detailed category-specific results, synthetic note snippets, and the annotation codebook.
[DOCX File (Microsoft Word File), 43 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Pseudocode for semantic similarity search pipeline.
[DOCX File (Microsoft Word File), 30 KB-Multimedia Appendix 2]

References

1. Sim JA, Huang X, Horan MR, et al. Natural language processing with machine learning methods to analyze unstructured patient-reported outcomes derived from electronic health records: a systematic review. Artif Intell Med. Dec 2023;146:102701. [doi: 10.1016/j.artmed.2023.102701] [Medline: 38042599]
2. Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB. Natural language processing in medicine: a review. Trends Anaesth Crit Care. Jun 2021;38:4-9. [doi: 10.1016/j.tacc.2021.02.007]
3. Wei Q, Franklin A, Cohen T, Xu H. Clinical text annotation - what factors are associated with the cost of time? AMIA Annu Symp Proc. 2018;2018:1552-1560. [Medline: 30815201]
4. Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263-1284. [doi: 10.1109/TKDE.2008.239]
5. Krishnakumar A. Active learning literature survey. University of California; 2007. URL: https://www.researchgate.net/publication/228971426 [Accessed 2026-08-26]
6. Settles B. Active learning literature survey. University of Wisconsin–Madison; 2009. URL: https://minds.wisc.edu/server/api/core/bitstreams/8a78cf83-0702-4157-aeff-3cacb62f9ad5/content [Accessed 2026-08-20]
7. Chen Y, Mani S. Active learning for unbalanced data in the challenge with multiple models and biasing. Presented at: Active Learning and Experimental Design Workshop in Conjunction with AISTATS 2010; May 26, 2010; Sardinia, Italy. URL: https://proceedings.mlr.press/v16/chen11a/chen11a.pdf [Accessed 2026-08-20]
8. Wei J, Zou K. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In: Inui K, JiangJ, NgV, WanX, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics; 2019:6382-6388. [doi: 10.18653/v1/D19-1670]
9. Li Z, Zhu H, Lu Z, Yin M. Synthetic data generation with large language models for text classification: potential and limitations. In: Bouamor H, Pino J, Bali K, editors. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2023:10443-10461. [doi: 10.18653/v1/2023.emnlp-main.647]
10. Feng SY, Gangal V, Wei J, et al. A survey of data augmentation approaches for NLP. In: ZongC, XiaF, Li W, NavigliR, editors. Findings of the Association for Computational Linguistics. Association for Computational Linguistics; 2021:968-988. [doi: 10.18653/v1/2021.findings-acl.84]
11. Şahin GG. To augment or not to augment? A comparative study on text augmentation techniques for low-resource NLP. Computational Linguistics. Apr 4, 2022;48(1):5-42. [doi: 10.1162/coli_a_00425]
12. Scroggins JK, Barcelona V, Hulchafo II, et al. Assessing the quality and performance of synthetic data augmentation to identify stigmatizing language in obstetric clinical notes. Nurs Outlook. 2026;74(3):102757. [doi: 10.1016/j.outlook.2026.102757] [Medline: 41962499]
13. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv. Preprint posted online on Jan 16, 2013. [doi: 10.48550/arXiv.1301.3781]
14. Salton G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley; 1989. ISBN: 978-0-201-12227-5
15. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics; 2019:3982-3992. [doi: 10.18653/v1/D19-1410]
16. Mahajan D, Poddar A, Liang JJ, et al. Identification of semantically similar sentences in clinical notes: iterative intermediate training using multi-task learning. JMIR Med Inform. Nov 27, 2020;8(11):e22508. [doi: 10.2196/22508] [Medline: 33245284]
17. Ormerod M, Martínez Del Rincón J, Devereux B. Predicting semantic similarity between clinical sentence pairs using transformer models: evaluation and representational analysis. JMIR Med Inform. May 26, 2021;9(5):e23099. [doi: 10.2196/23099] [Medline: 34037527]

18. Lopez I, Swaminathan A, Vedula K, et al. Clinical entity augmented retrieval for clinical information extraction. NPJ Digit Med. Jan 19, 2025;8(1):45. [doi: 10.1038/s41746-024-01377-1] [Medline: 39828800]
19. Arzideh K, Schäfer H, Idrissi-Yaghir A, et al. Improving retrieval augmented generation for health care by fine-tuning clinical embedding models: development and evaluation study. J Med Internet Res. Mar 25, 2026;28:e82997. [doi: 10.2196/82997] [Medline: 41880603]
20. Zhang B, Rouhizadeh H, Chen Y, et al. Retrieval-augmented generation in biomedicine: a survey of technologies, datasets, and clinical applications. Research Square. Preprint posted online on Dec 15, 2025. [doi: 10.21203/rs.3.rs-8330917/v1]
21. Shattell MM. Stigmatizing language with unintended meanings: “persons with mental illness” or “mentally ill persons”? Issues Ment Health Nurs. Mar 2009;30(3):199. [doi: 10.1080/01612840802694668] [Medline: 19291498]
22. Sun M, Oliwa T, Peek ME, Tung EL. Negative patient descriptors: documenting racial bias in the electronic health record. Health Aff (Millwood). Feb 2022;41(2):203-211. [doi: 10.1377/hlthaff.2021.01423] [Medline: 35044842]
23. Benkert R, Cuevas A, Thompson HS, Dove-Meadows E, Knuckles D. Ubiquitous yet unclear: a systematic review of medical mistrust. Behav Med. 2019;45(2):86-101. [doi: 10.1080/08964289.2019.1588220] [Medline: 31343961]
24. Martin JA, Hamilton BE, Osterman MJK. Births in the United States, 2023. National Center for Health Statistics; 2022. URL: https://www.cdc.gov/nchs/data/databriefs/db507.pdf [Accessed 2026-08-20]
25. Himmelstein G, Bates D, Zhou L. Examination of stigmatizing language in the electronic health record. JAMA Netw Open. Jan 4, 2022;5(1):e2144967. [doi: 10.1001/jamanetworkopen.2021.44967] [Medline: 35084481]
26. Barcelona V, Scharp D, Idnay BR, et al. A qualitative analysis of stigmatizing language in birth admission clinical notes. Nurs Inq. Jul 2023;30(3):37073504. [doi: 10.1111/nin.12557] [Medline: 37073504]
27. Barcelona V, Horton RL, Rivlin K, et al. The power of language in hospital care for pregnant and birthing people: a vision for change. Obstet Gynecol. Oct 1, 2023;142(4):795-803. [doi: 10.1097/AOG.0000000000005333] [Medline: 37678895]
28. Barcelona V, Scharp D, Moen H, et al. Using natural language processing to identify stigmatizing language in labor and birth clinical notes. Matern Child Health J. Mar 2024;28(3):578-586. [doi: 10.1007/s10995-023-03857-4] [Medline: 38147277]
29. Barcelona V, Scroggins JK, Scharp D, et al. Secondary qualitative analysis of stigmatizing and nonstigmatizing language used in hospital birth settings. J Obstet Gynecol Neonatal Nurs. Jan 2025;54(1):112-122. [doi: 10.1016/j.jogn.2024.10.003] [Medline: 39577837]
30. Park J, Saha S, Chee B, Taylor J, Beach MC. Physician use of stigmatizing language in patient medical records. JAMA Netw Open. Jul 1, 2021;4(7):e2117052. [doi: 10.1001/jamanetworkopen.2021.17052] [Medline: 34259849]
31. Scroggins JK, Hulchafo II, Harkins S, et al. Identifying stigmatizing and positive/preferred language in obstetric clinical notes using natural language processing. J Am Med Inform Assoc. Feb 1, 2025;32(2):308-317. [doi: 10.1093/jamia/ocae290] [Medline: 39569431]
32. Sentence-transformers/multi-qa-distilbert-cos-v1. Hugging Face. 2024. URL: https://huggingface.co/sentence-transformers/multi-qa-distilbert-cos-v1 [Accessed 2026-08-20]
33. Excoffier JB, Roehr T, Figueroa A, Papaioannou JM, Bressem K, Ortala M. Generalist embedding models are better at short-context clinical semantic search than specialized embedding models. arXiv. Preprint posted online on Jan 3, 2024. [doi: 10.48550/arXiv.2401.01943]
34. Johnson J, Douze M, Jegou H. Billion-scale similarity search with GPUs. IEEE Trans Big Data. 2021;7(3):535-547. [doi: 10.1109/TBDATA.2019.2921572]
35. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press; 2009. URL: https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf [Accessed 2026-08-20]
36. Lingren T, Deleger L, Molnar K, et al. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. J Am Med Inform Assoc. 2014;21(3):406-413. [doi: 10.1136/amiajnl-2013-001837] [Medline: 24001514]
37. South BR, Mowery D, Suo Y, et al. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. J Biomed Inform. Aug 2014;50:162-172. [doi: 10.1016/j.jbi.2014.05.002] [Medline: 24859155]
38. Pangakis N, Wolken S. Keeping humans in the loop: human-centered automated annotation with generative AI. ICWSM. 2025;19:1471-1492. [doi: 10.1609/icwsm.v19i1.35883]
39. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Oct 25-29, 2014; Doha, Qatar. [doi: 10.3115/v1/D14-1162]

40. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on Oct 11, 2018. [doi: 10.48550/arXiv.1810.04805]
41. Arora S, Liang Y, Ma T. A simple but tough-to-beat baseline for sentence embeddings. Presented at: 5th International Conference on Learning Representations (ICLR 2017); Apr 24-26, 2017; Toulon, France. URL: https://openreview.net/pdf?id=SyK00v5xx [Accessed 2026-08-20]
42. Ethayarajh K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP. Association for Computational Linguistics; 2019:55-65. [doi: 10.18653/v1/D19-1006]
43. Kholghi M, Sitbon L, Zuccon G, Nguyen A. Active learning reduces annotation time for clinical concept extraction. Int J Med Inform. Oct 2017;106:25-31. [doi: 10.1016/j.ijmedinf.2017.08.001] [Medline: 28870380]
44. Wei Q, Chen Y, Salimi M, et al. Cost-aware active learning for named entity recognition in clinical text. J Am Med Inform Assoc. Nov 1, 2019;26(11):1314-1322. [doi: 10.1093/jamia/ocz102] [Medline: 31294792]
45. Nachtegael C, De Stefani J, Lenaerts T. A study of deep active learning methods to reduce labelling efforts in biomedical relation extraction. PLoS ONE. 2023;18(12):e0292356. [doi: 10.1371/journal.pone.0292356] [Medline: 38100453]
46. Alsentzer E, Murphy J, Boag W, et al. Publicly available clinical. In: Rumshisky A, Roberts K, Bethard S, Naumann T, editors. Proceedings of the 2nd Clinical Natural Language Processing Workshop. Association for Computational Linguistics; 2019:72-78. [doi: 10.18653/v1/W19-1909]
47. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. Jul 2009;45(4):427-437. [doi: 10.1016/j.ipm.2009.03.002]
48. Ward N, Marco D. A collection of pragmatic-similarity judgments over spoken dialog utterances. In: Calzolari N, Kan MY, Hoste V, Lenci A, Sakti S, Xue N, editors. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING). ELRA and ICCL; 154-163. [doi: 10.63317/5iqq22vfx8n2]

Abbreviations

BERT: bidirectional encoder representations from transformers
EHR: electronic health record
IRB: institutional review board
NLP: natural language processing

Edited by Arriel Benis; peer-reviewed by Ama Quansah, Dillon Chrimes, Mohammad Al-Agil; submitted 30.Sep.2025; final revised version received 13.Aug.2026; accepted 13.Aug.2026; published 22.Sep.2026

Please cite as:
Scroggins JK, Hulchafo II, Barcelona V, Davoudi A, Moen H, Harkins S, Scharp D, Cato K, Tadiello M, Topaz M
Semantic Similarity Search Approach to Extract Exemplars of Stigmatizing and Positive Language in Obstetric Clinical Notes: Exploratory Study
JMIR Med Inform 2026;14:e85088
URL: https://medinform.jmir.org/2026/1/e85088
doi: 10.2196/85088