Positive Mining from LLM Seeds: A Semi-Supervised Graph Based Approach to Train Rare Event Classifiers

Sasan Tavakkol
Google Research
New York, NY, USA

Lin Chen
Google Research
New York, NY, USA

Max Springer
University of Maryland
College Park, MD, USA

Abigail Schantz
Google Research
New York, NY, USA

Blaž Bratanič
Google Research
New York, NY, USA

Vincent Cohen-Addad
Google Research
New York, NY, USA

MohammadHossein Bateni
Google Research
New York, NY, USA

Abstract

Scarcity of labeled data, especially for rare events, hinders training effective machine learning models. This paper proposes SYNAPSE-G (Synthetic Augmentation for Positive Sampling via Expansion on Graphs), a novel pipeline leveraging Large Language Models (LLMs) to generate synthetic training data for rare event classification, addressing the cold-start problem. SYNAPSE-G generates synthetic rare event examples using an LLM, which then serve as seeds for semi-supervised label propagation on a similarity graph constructed between the seeds and a large unlabeled dataset. This identifies candidate positive examples, subsequently labeled by an oracle (human or LLM). The expanded dataset then trains/fine-tunes a classifier. We theoretically analyze how the quality (validity and diversity) of the synthetic data impacts the precision and recall of our method. Experiments on the imbalanced SST2 dataset demonstrate SYNAPSE-G's effectiveness in finding positive labels, outperforming baselines including nearest neighbor search. We use publicly available synthetic data to focus on evaluating our method's efficacy.

CCS Concepts

• **Computing methodologies** → **Supervised learning**; **Learning paradigms**; **Unsupervised learning**.

Keywords

Machine learning, learning paradigms, graph-based sampling, supervised learning

ACM Reference Format:
Sasan Tavakkol, Lin Chen, Max Springer, Abigail Schantz, Blaž Bratanič, Vincent Cohen-Addad, and MohammadHossein Bateni. 2025. Positive Mining from LLM Seeds: A Semi-Supervised Graph Based Approach to Train Rare Event Classifiers. In *Proceedings of Machine Learning on Graphs in the Era of Generative Artificial Intelligence (Proceedings of MLoG-GenAI Workshop (KDD '25))*. ACM, New York, NY, USA, 9 pages.

Proceedings of MLoG-GenAI Workshop (KDD '25), Toronto, ON, Canada

1 Introduction

The rapid emergence of new trends on social media and the internet, such as misinformation [29], fraud [10], and hate speech [20, 26], necessitates the development of effective classifiers for timely detection and mitigation [2]. However, the dynamic nature and novelty of these trends often results in scarcity of labeled data for training supervised models [25]. This research tackles this "cold-start" problem by introducing a pipeline that leverages LLMs and semi-supervised learning for collecting labeled data for training robust classifiers. Our method, SYNAPSE-G, offers a practical and generalizable solution for addressing emerging online threats.

SYNAPSE-G augments real labeled data with LLM-generated synthetic data in three stages: (1) **Synthetic Data Generation:** An LLM generates synthetic rare event examples (e.g., hate speech), creating a seed set. (2) **Label Propagation:** This seed set is used in semi-supervised label propagation, expanding the labeled set by connecting seeds to similar unlabeled instances on a similarity graph. (3) **LLM-Based Refinement (Optional):** An LLM rater can refine propagated labels, mitigating errors.[1] The augmented dataset then trains/fine-tunes a classifier. This paper makes the following key contributions:

- We propose SYNAPSE-G, a novel pipeline combining LLM-based synthetic data generation with graph-based semisupervised learning for rare event classification.
- We provide a theoretical analysis of how the quality of the synthetic data (validity and diversity) influences the precision and recall of the label propagation process.
- We empirically demonstrate the effectiveness of SYNAPSE-G on an artificially imbalanced SST2 dataset, showing significant performance improvements over baselines.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 formally defines the problem, Section 4 details our proposed method (SYNAPSE-G), Section 5 presents

[1] Human annotators can also be used.

a theoretical analysis, Section 6 discusses experimental results, and 7 concludes.

2 Related Work

Our work relates to several areas: retrieval, diversity sampling, positive mining, and label propagation on graphs.

Retrieval. Retrieval methods aim to identify relevant items from a large dataset based on a given query. Traditional approaches rely on lexical matching techniques such as BM25 [23], which score documents based on term frequency and inverse document frequency (TF-IDF). While effective for keyword-based search, these methods fail to capture semantic meaning limiting their performance on more complex retrieval tasks. To overcome these limitations, dense retriever methods have emerged, leveraging neural embeddings to map both queries and documents into a shared vector space where relevance can be measured with standard similarity metrics [13, 14, 18, 33].

Closely related to our work is that of Hypothetical Document Embeddings (HyDE) [8], which bypasses the need for relevance labels. This method generates a synthetic document which is then used to retrieve similar (real) documents from a dataset, allowing for effective search without fine-tuning or task-specific supervision.

Diversity Sampling. Obtaining high-quality data from humans can be challenging or even impractical due to high costs and privacy concerns [16]. Several studies have further showcased that human-generated data, being inherently prone to biases or errors, may not even be ideal for model training on all tasks in general [9, 12, 27]. In mitigating these issues, a burgeoning area of research has explored the task of *generating* data which more diversely samples the training space [7, 19]. In the current problem, data samples of interest are rare and a simpler randomized selection of data points will struggle to recover a diverse set of examples which encompass the rare event. As such, to carefully pull out a diverse set of (rare) positive examples, we leverage synthetic data to guide our graph theoretic approaches towards the known small subset.

Positive Mining. Positive (or negative) mining seeks to identify instances that are likely to belong to a target (our non-target) class, often used to help refine decision boundaries in the data space especially when labeled data is scarce or imbalanced [21]. Traditional approaches, such as hard negative mining in contrastive learning [33], select negatives that are close to the decision boundary to improve model generalization [11]. In our setting, positive mining is at the core of detecting rare events within a large unlabeled dataset.

Label Propagation on Graphs. Our work falls in the domain of "label propagation", a fundamental approach in semi-supervised learning that leverages the structure of data to infer labels for unlabeled instances. This method assumes that similar points should have similar labels, enforcing smoothness in the label distribution [3]. However, label propagation relies on the presence of an initial labeled dataset and assumes that the underlying graph structure accurately captures class boundaries [1, 22, 31].

Algorithm 1 Active Learning Framework for Rare Event Detection

Require: Unlabeled dataset $\mathcal{D}_U$, iterations T, batch size B

1: Initialize $i \leftarrow 1$
2: **while** $i \leq T$ **do**
3: Select a batch $\mathcal{B}_i \subset \mathcal{D}_U$ of size B using selection strategy (based on the current state of the algorithm).
4: Obtain labels $\mathcal{L}_i = \{(x, y) | x \in \mathcal{B}_i\}$ from the oracle (human labelers or an LLM rater).
5: Update the algorithm's internal model based on $\mathcal{L}_i$ and potentially previous labeled sets $\bigcup_{j=1}^{i-1} \mathcal{L}_j$.
6: $i \leftarrow i + 1$
7: **end while**

In contrast, our approach addresses a fundamentally different problem: identifying rare, positive, instances within a large *unlabeled* dataset without an existing set of labeled examples. Rather than relying on label propagation from known labels, we generate synthetic instances for the rare event and use their embedding to identify similar real instances. This removes the need for model training or iterative graph-based updates. This distinction makes our approach particularly suitable for applications where positive instances are extremely rare and must be identified without prior ground truth labels.

3 Preliminaries & Problem Definition

The detection of rare events amidst a vast expanse of routine occurrences is a critical task across a multitude of real-world domains. From identifying fraudulent transactions in financial systems to pinpointing equipment malfunctions in industrial settings and diagnosing rare diseases in healthcare, these infrequent yet impactful events demand accurate and timely identification. This work tackles the problem of binary classification in such domains, where the "rare event" class is significantly underrepresented.

Formally, let $\mathcal{D}$ denote the data domain. Each observation is represented by a feature vector x, and the data distribution is denoted by $\Pr(x, y)$, where $y \in \{0, 1\}$ is the class label ($y = 1$ is the rare event).

Our setting departs from traditional supervised learning. Instead of a readily available labeled dataset, we confront a **complete absence of labeled data** initially. We operate within an active learning framework, tailored for iterative label acquisition. The algorithm begins with a completely unlabeled dataset, denoted by $\mathcal{D}_U = \{x_j\}_{j=1}^{n_U}$. The full active learning process, including detailed descriptions of selection strategies and model updates, is presented in Algorithm 1.

The objective is to maximize the cumulative precision and recall across all queried batches up to each step i. Formally, at each step i, let P_i and R_i represent the precision and recall, respectively, calculated over the union of all labeled sets acquired up to that point: $\bigcup_{j=1}^{i} \mathcal{L}_j$. The algorithm aims to maximize both P_i and R_i for all $i \in \{1, ..., T\}$ as we increase the *ratio of queried data*, or the fraction of data we obtain labels from through the oracle, $\frac{|\cup_{j=1}^{i} \mathcal{L}_j|}{|\mathcal{D}_U|}$. This reflects the goal of efficiently identifying as many rare events as possible with minimal false positives. The core challenges are the **cold start** (no initial labeled data) and the **severe class imbalance**.

4 Methodology

Our method addresses rare event classification with limited labeled data by generating and leveraging synthetic data. The core is a three-stage pipeline integrating LLMs with semi-supervised learning to augment a small (or non-existent) initial set of labeled real data. Figure 1 provides an overview.

4.1 Synthetic Data Generation (Seed Set Creation)

To address the cold start problem (absence of initial labeled data), we use an LLM to generate an initial seed set of labeled data, $\mathcal{D}_S$. This synthetic dataset bootstraps the learning process. In practice, one should use carefully crafted prompts to guide the LLM towards generating examples representative of the rare event class. However, we omit this step here as it is outside the scope of this research, and instead focus on selecting the best data points from a pool of synthetically generated data. We will compare two selection methods in the experiments: random sampling and the Adaptive Coverage Sampling (ACS) approach of [32] which selects k points that collectively cover a c-portion of the dataset, maximizing diversity within the synthetic data.

4.2 Label Propagation to Unlabeled Data

This stage expands the labeled dataset by propagating labels from the selected synthetic seed data ($\mathcal{D}_S$) to the unlabeled data ($\mathcal{D}_U$) via semi-supervised learning. We assume access to a large corpus of unlabeled data, $\mathcal{D}_U$, representative of the target domain and containing both rare event and non-event instances. Both synthetic and unlabeled data are transformed into numerical embeddings (e.g., using BERT [5] or Gecko [17]) so that semantically similar data points are close in the embedding space. We propose two semi-supervised propagation approaches which we denote as *Iterative Bipartite Graph (IBG)* and *Graph-Based Label Expansion (GBLE)*.

Iterative Bipartite Graph (IBG). IBG iteratively refines a bipartite graph between known positives (V_P, initially $\mathcal{D}_S$) and unlabeled data (V_U). Edges are created based on cosine similarity exceeding a threshold, then pruned to retain only the top d_{max} connections per node in V_P. Connected V_U nodes are queried for labels. New positives are added to V_P, and the process repeats. See Algorithm 2 for details.

Graph-Based Label Expansion (GBLE). This approach utilizes a more global graph structure in conjunction with a standard propagation technique [22, 34]. The specific algorithm is detailed Algorithm 3. In brief, a similarity graph is constructed over the entire real dataset and the initial synthetic seed data. The known labels (positive and negative) are then propagated through this graph. The algorithm assigns a learned weight to each real data point (its likelihood of being positive) based on a convex objective function for propagation that is stemming from a multi-class generalization of the quadratic cost criterion of [3] and [30] for structured prediction, and the top K points are selected as candidates for labeling. K is dynamically adjusted each iteration: $K = K_0/p_{prev}$, where p_{prev} is the precision from the previous iteration, aiming to find approximately K_0 new positives per round.

Algorithm 2 Iterative Bipartite Graph (IBG)

Require: Unlabeled data $\mathcal{D}_U$, Initial positive seeds $\mathcal{D}_S$, Similarity threshold τ, Maximum degree d_{max}, Number of iterations T.
Ensure: Labeled data $\mathcal{L}$.
Initialize $V_P = \mathcal{D}_S$, $\mathcal{L} = \mathcal{D}_S$.
for $t = 1$ to T **do**
  $V_U = \mathcal{D}_U$ // Remaining unlabeled data
  Construct bipartite graph $G_B = (V_P, V_U, E_B)$.
  for $v_i \in V_P$ **do**
    for $v_j \in V_U$ **do**
      Calculate cosine similarity $sim(v_i, v_j)$.
      if $sim(v_i, v_j) > \tau$ **then**
        Add edge e_{ij} to E_B.
      end if
    end for
    Sort neighbors of v_i in V_U by similarity (descending).
    Keep only top d_{max} neighbors, removing other edges from E_B.
  end for
  $\mathcal{B}_t \leftarrow$ Nodes in V_U connected to V_P in G_B.
  Obtain labels $\mathcal{L}_t$ for $\mathcal{B}_t$ from the oracle.
  $V_P \leftarrow V_P \cup \{v \in \mathcal{B}_t \mid \text{label}(v) = \text{positive}\}$. // Add new positives
  $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{L}_t$
  $\mathcal{D}_U \leftarrow \mathcal{D}_U \setminus \mathcal{B}_t$
end for
return $\mathcal{L}$

Algorithm 3 Graph-Based Label Expansion

Require: Similarity graph $G = (V, E)$, initial labels Y_0 (partially labeled), iterations T.
Ensure: Final label assignments Y_T.
Initialize $Y^{(0)} = Y_0$.
for $t = 1$ to T **do**
  Construct the normalized adjacency matrix W from G:

$$W_{ij} = \begin{cases} \frac{1}{\deg(v_i)} & \text{if } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases}$$

  Propagate labels: $Y^{(t)} = WY^{(t-1)}$.
  Reinforce initial labels: For all nodes v_i with initial labels in Y_0, set $Y_i^{(t)} = Y_{0,i}$.
end for
return $Y^{(T)}$.

5 Theoretical Analysis

To understand how prompt quality (validity and diversity of synthesized data) impacts algorithm performance, we analyze a simplified, single-iteration version of our algorithm. We model data and relationships using an undirected, simple, d-regular graph $G = (V, E)$. Each node $v \in V$ is a data point with a binary label $y(v) \in \{0, 1\}$ (1: positive, 0: negative). Let $S \subseteq V$ represent the LLM-generated synthesized data. The algorithm queries S and then the neighbors

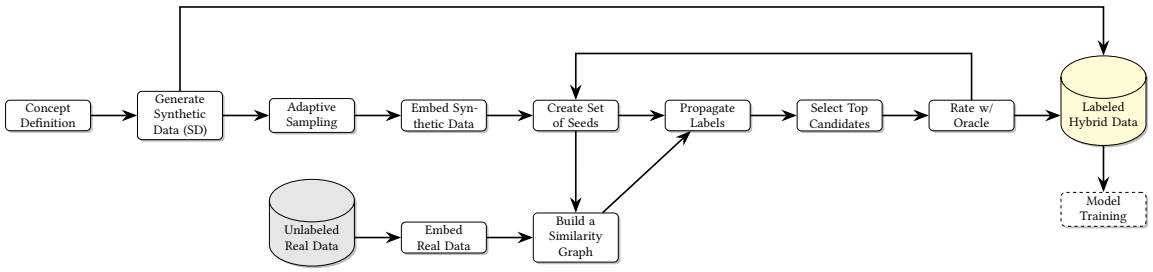


Figure 1: Overview of the SYNAPSE-G pipeline for rare event classification. The pipeline integrates synthetic data generation (top branch) with real data processing (bottom branch). LLM-generated synthetic data, after adaptive sampling and embedding, forms a set of positive seeds. Unlabeled real data is also embedded. A similarity graph connects seeds and real data, enabling label propagation to identify top candidates. An oracle rates these candidates, creating a labeled hybrid dataset for model training. The iterative nature is shown by the feedback loop from "Rate w/ Oracle" to "Create Set of Seeds".

of positive seeds, $N(S_+)$. In proving theoretical results on our algorithms expected guarantees, we first invoke a few careful assumptions on the input which hold in practice.

Assumption 1 (Diversity of Synthesized Data)**.** S exhibits two properties. (1) *Independence*: S forms an independent set on the graph. (2) *Limited Overlap*: No vertex in V is adjacent to more than two vertices in S.

Define $N(v) \coloneqq \{u \in V \mid (u,v) \in E\} \cup \{v\}$ and $N(S) \coloneqq \bigcup_{v \in S} N(v)$. The *vertex expansion ratio* is $h(S) \coloneqq \frac{|N(S)|}{|S|} \geq 1$, quantifying the *coverage* of S.

Assumption 2 (Partition and Proportion of Positive Examples)**.** S is partitioned into $S_+ = \{v \in S \mid y(v) = 1\}$ and $S_- = \{v \in S \mid y(v) = 0\}$. Thus, we define $p = \frac{|S_+|}{|S|}$ is the proportion of positive examples (validity).

Assumption 3 (Labeling Probabilities)**.** If $u \in V$ is adjacent to exactly n positive vertices in S, then denote the probability that u is labeled by q_n for $n = 1, 2$. Assume $0 < q_1 < q_2 < 2q_1$.[2]

Stemming from these assumptions, we obtain the following proposition with the proof deferred to Appendix A due to space constraints.

Proposition 1. *Let $Q \coloneqq S \cup N(S_+)$ denote the queried vertices and P be the number of positive examples in Q. Then,*

$$\mathbb{E}\left[\frac{P}{|Q|} \mid S\right] = (2q_1 - q_2) + \frac{1 + q_2\left(d + \frac{1}{p}\right) - q_1\left(d + \frac{2}{p}\right)}{\frac{1-p}{p} + h\left(S_+\right)}$$

$$\mathbb{E}\left[\frac{P}{|V|} \mid S\right] = \frac{p|S|}{|V|}\left((1 - 2q_1 + q_2) + (q_2 - q_1)d + (2q_1 - q_2)h(S_+)\right)$$

where $\frac{P}{|Q|}$ is the precision and $\frac{P}{|V|}$ is the recall.

This result explores how two key dimensions of synthesized data quality – *validity* (p) and *diversity* ($h(S_+)$) – impact precision and recall. Intuitively, this first equality proves that recall increases with both p (higher probability of synthesized seeds being truly positive) and $h(S_+)$ (greater diversity, allowing exploration of more examples). This aligns with intuition. The relationship between precision and diversity, however, is more nuanced. Precision always increases with p, as expected, but the impact of diversity ($h(S_+)$) on precision depends on the magnitude of p relative to a threshold determined by q_1, q_2, and d. This threshold, $\frac{2q_1 - q_2}{1+(q_2-q_1)d}$, is increasing in q_1 and decreasing in q_2. In segregating the results based on this thresholding, we can conclude the following important facts:

- **High Validity Regime** $\left(p > \frac{2q_1 - q_2}{1+(q_2-q_1)d}\right)$**:** When the validity of the synthesized positives is sufficiently high (large p, small q_1, and/or large q_2), precision *decreases* with increasing diversity. Intuitively, if neighbors of single positive seeds are unlikely to be positive (low q_1), then maximizing precision requires focusing on regions with *overlapping* neighborhoods (lower $h(S_+)$), increasing the chance of finding nodes adjacent to *multiple* positive seeds (higher q_2).
- **Low Validity Regime** $\left(p < \frac{2q_1 - q_2}{1+(q_2-q_1)d}\right)$**:** When the validity is low (small p, large q_1, and/or small q_2), precision *increases* with diversity. In this case, even nodes adjacent to a *single* positive seed are sufficiently likely to be positive, making greater coverage (higher $h(S_+)$) beneficial for precision.

This analysis reveals a crucial interplay between the validity and diversity of synthesized data and their combined effect on precision, offering valuable insights for designing effective prompt engineering and data selection strategies.

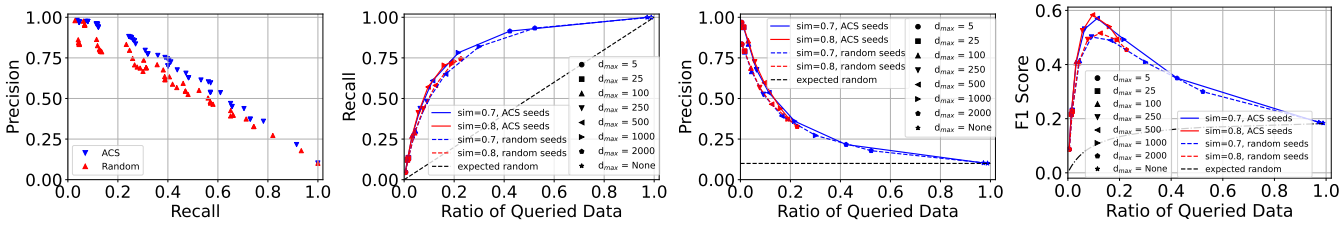
6 Experimental Results

We here validate our methods on two representative datasets: the Stanford Sentiment Treebank 2 (SST2 [28]) and Measuring Hate Speech (MHS [15, 24]). We proceed to define the experimental framework for each dataset and results which demonstrate the considerable improvement of SYNAPSE-G.

6.1 SST2 Dataset

The SST2 dataset is a standard sentiment analysis benchmark comprised of movie review sentences with positive/negative labels. We augment this data with a public synthetic SST2 dataset generated using GPT [6].

[2]Justification: Each link from a positive example independently assigns a positive label to its neighbor with probability q_1. Then $q_2 = 1 - (1 - q_1)^2 = 2q_1 - q_1^2$, satisfying $q_1 < q_2 < 2q_1$.



(a) Precision vs. Recall **(b) Recall vs. Query Ratio.** **(c) Precision vs. Query Ratio.** **(d) F1 Score vs. Query Ratio.**

Figure 2: Experimental results on the imbalanced SST2 dataset. (a) Precision vs. Recall, comparing ACS and random seed selection. (b) Recall vs. Query Ratio, showing the benefit of ACS seeds. (c) Precision vs. Query Ratio, illustrating the impact of parameters. (d) F1 Score vs. Query Ratio, showing the impact of parameters. ACS consistently outperforms random seed selection across all metrics. Figures (b), (c), and (d) demonstrate the impact of varying similarity thresholds and maximum degree constraints. Curves above the expected random performance (diagonal in (b) and horizontal/curve in (c)/(d) respectively) indicate a benefit of the graph-based approach.

To simulate a rare event, we create a class-imbalanced SST2 training set, keeping all negative examples and subsampling positive examples to 10% of the modified set. We select 100 positive synthetic examples as seeds. Table 1 summarizes dataset statistics.

We focus on *single-shot* (one iteration) and *iterative* evaluations, using the imbalanced SST2 "train" split. The 100 positive seeds are selected from the synthetic dataset (randomly or via ACS [32] with coverage $c = 0.5$). We construct a bipartite graph connecting seed and training examples, using cosine similarity of pre-trained Gecko embeddings [17] for edge creation.

Baselines. We compare our approach against the following methods:

- **Random Selection (Theoretical):** expected values, calculated analytically
- **Random Seeds + Bipartite Graph:** 100 random positive seeds; bipartite graph constructed as above and connected real data points are labeled
- **ACS Seeds + Bipartite Graph:** Identical to (2), but using ACS for seed selection [32].
- **Graph Based Label Expansion:** Similarity graph on the *entire* real dataset + 100 initial ACS seeds. GBLE (Algorithm 3) propagates labels. Top K points (highest positive weights) are candidates; $K = 100/p_{prev}$ (p_{prev}: previous iteration's precision).

ACS and GBLE are included as baselines as they utilize data point relationships.

Results and Discussion. We evaluate performance using precision-recall curves, recall vs. query ratio, precision vs. query ratio, and F1 score vs. query ratio, analyzing the impact of similarity threshold and maximum degree (d_{max}).

Dataset	All	Pos.	Neg.
Original SST2 Train	67349	37569	29780
Original Synthetic	5000	2488	2512
Imbalanced SST2 Train	33088	3308	29780
Synthetic Seeds (Positive)	100	100	0

Table 1: Dataset statistics for SST2.

Figure 2a shows that ACS seed selection consistently achieves higher precision for any given recall compared to random seed selection. This highlights the benefit of diverse and representative seed sets.

Figure 2 depicts the recall, precision, and F1 score vs. query ratio. Varying similarity thresholds (0.7, 0.8) and d_{max} are analyzed.

Figure 2b plots the recall versus query ratio. The dashed line depicts expected recall from random performance. All methods here achieve a recall of 1.0 when querying all data. Crucially, ACS seeds consistently achieve higher recall than random seeds for a given query ratio. We note that a higher d_{max} allows more connections in the bipartite graph, generally increasing recall, but with diminishing returns. A higher similarity threshold (0.8) generally results in better performance but limits reachability and consequently the recall.

Figure 2c show the precision versus query ratio, with the dashed line representing the base positive rate (10%). Both graph-based methods (ACS and random seeds) achieve significantly higher precision than random selection, particularly at low query ratios. ACS consistently outperforms random seed selection. Higher similarity thresholds and increasing d_{max} (up to a point) generally improve precision.

Lastly, Figure 2d shows the F1 score, with the black line representing expected random performance. The F1 score, balancing precision and recall, initially increases with the query ratio, peaks, and then decreases. ACS-selected seeds generally outperform random seeds. A higher similarity threshold leads to higher peak F1 scores. Increasing d_{max} initially improves the F1 score, with diminishing returns and potential slight performance decreases at very high d_{max} and high query ratios. The best F1 scores are below 0.6, highlighting the difficulty of the rare event detection problem.

As we see in the above, ACS consistently outperforms random seed selection. Crucially, higher similarity thresholds generally improve performance, especially at lower query ratios. Additionally, increasing d_{max} improves performance up to a point, with diminishing returns. These results highlight the importance of strategic seed selection (ACS) and parameter tuning.

Figure 3 compares two iterative strategies, Iterative Bipartite Graph (IBG) and Graph-Based Label Expansion (GBLE), within an "ideal" scenario (precision = 1). GBLE significantly outperforms IBG, achieving much higher recall for a given query ratio. Notably, IBG plateaus quickly while GBLE leverages the full graph structure and both positive/negative labels.

6.2 MHS Dataset

To further evaluate SYNAPSE-G on a naturally occurring rare event task with real-world relevance, we utilized the MHS dataset [15, 24]. This dataset contains 39,565 comments with annotations from 7,912 annotators (135,556 total rows).

Whereas the SST2 dataset is more straightforward in its labeling, the MHS dataset is more complex and relies on further preprocessing to define a specific rare event. Specifically, the dataset comprises social media comments (YouTube, Reddit, Twitter) annotated across 10 ordinal labels to derive a continuous "hate speech score", with each sentence also being labeled according to the specific groups or demographics targeted. As such, we here define a specific rare event binary label for MHS: hate speech targeting transgender individuals. Concretely, we create a label which is positive if any annotator marked a comment as targeting any of the sub-categories: transgender men, transgender women, or unspecified transgender individuals. We do this by applying a logical OR across the three noted subcategory annotations. This resulted in 2,598 comments (6.5%) being labeled positive, representing an organically imbalanced dataset relevant to real-world challenges.

For the cold-start scenario, we used 1,000 LLM (Llama-2) generated comments related to the dataset's topics, sourced from [4]. Within this synthetic set, only 19 comments were positive for our *target_gender_transgender* label. These 19 synthetic examples served as the initial positive seeds for SYNAPSE-G to identify real positive instances in the unlabeled data pool. Consistent with our SST2 experiments, we used pre-trained Gecko embeddings to represent comments for constructing the similarity graph.

Baselines. We further design a practical baseline, "LR-Baseline", which adopts an iterative active learning approach using a simple classifier. We establish the baseline using a logistic regression model (initialized with `scikit-learn`'s default parameters) trained on an initial set comprising 19 positive synthetic seeds augmented with 19 randomly sampled known negative instances from the dataset. These examples are represented using pre-trained Gecko embeddings. The iterative refinement process then proceeds as follows: In each iteration, a subset of unlabeled data points, constrained by an inference budget (B) to ensure practical feasibility by avoiding inference over the entire dataset, is selected. The current logistic regression model predicts positivity probabilities for this subset,

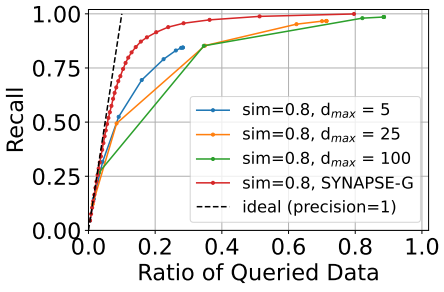


Figure 3: Recall vs. Ratio of Queried Data for *iterative* rare event detection on the imbalanced SST2 dataset. "Ideal" is perfect precision. The graph-Based Label Expansion (GBLE) significantly outperforms Iterative Bipartite Graph (IBG).

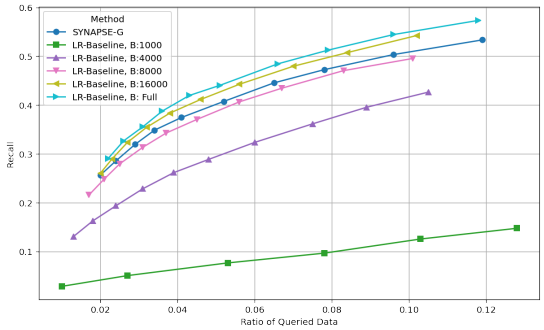


Figure 4: Recall vs. Ratio of Queried Data for *iterative* rare event detection on the imbalanced MHS dataset.

and the top K candidates with the highest predicted probabilities are chosen for labeling via an oracle (K is dynamically adjusted as $K = K_0/p_{prev}$, where $K_0 = 100$, consistent with SYNAPSE-G). These newly labeled instances are then incorporated into the training set, and the logistic regression model is retrained. The inference budget (B) is a critical parameter for maintaining the practicality of the approach.

Results and Discussion. Figure 4 summarizes SYNAPSE-G's recall performance using the default parameters ($K_0 = 100$) and compares it to LR-Baseline's recall for different Inference Budgets ($B = [1000, 4000, 8000, 16000]$) and without a budget ($B = 39,565$). Crucially, both SYNAPSE-G (via graph construction) and LR-Baseline (as input features) leverage the same Gecko embedding space, ensuring a fair comparison in terms of input features.

It is important to note, however, that the LR-Baseline was initialized with access to 19 known true negative labels in addition to the 19 synthetic positive seeds. This provides the LR-Baseline with an information advantage compared to SYNAPSE-G's strict cold-start setting, which assumes access only to synthetic positives and unlabeled data. Despite this initial advantage for LR, comparing their performance (see Figure 4) reveals a clear trade-off between computational budget and recall. SYNAPSE-G remains highly efficient at discovering rare positive instances, achieving substantial recall with minimal labeling effort. For example, by labeling only 2.4% of the data, SYNAPSE-G successfully identifies 28.6% of the true positive comments. Increasing the labeling budget to just 5% allows SYNAPSE-G to retrieve 40.8% of the positives. In contrast, LR-Baseline requires a large inference budget ($B = 8000$, inferring on 20% of the data per round) to reach similar recalls. Furthermore, only with very large or unlimited budgets ($B = 16k$ or Full), involving significant computational cost per iteration, does LR-Baseline outperform SYNAPSE-G in terms of recall.

Dataset	All	Pos.	Neg.
MHS Train	39565	2598	36967
Synthetic	1000	19	981
Synthetic Seeds (Positive)	19	19	0

Table 2: Dataset statistics for MHS.

This comparison highlights SYNAPSE-G's practical advantages for real-world large-scale applications involving extreme rarity. While the current dataset exhibits moderate imbalance (6.5% positive), real-world scenarios often present much more severe challenges (e.g., identifying a few thousand target posts among billions). In such extreme cases, the LR-baseline's need to infer over massive subsets (large B) or the entire dataset to find a handful of positives becomes computationally infeasible, potentially leading to complete failure. SYNAPSE-G, however, is designed to handle such scenarios more effectively. By leveraging the graph structure to focus exploration around known positive seeds (synthetic or newly discovered real ones) and their neighbors, it can maintain a high action rate (precision among queried candidates) even when positives are extremely sparse. This targeted approach makes SYNAPSE-G a significantly more scalable and practical solution for discovering truly rare events in massive datasets where methods requiring broad dataset inference at each step are not viable.

7 Conclusion

In conclusion, SYNAPSE-G, our proposed framework leveraging synthetic data generation and graph-based semi-supervised learning, offers a compelling approach to the challenging task of rare event classification. Our theoretical underpinnings illuminate the critical balance between the fidelity and diversity of the synthesized data, providing insights into the method's efficacy. Empirical evaluations on benchmark datasets demonstrate the practical effectiveness of SYNAPSE-G, showcasing its superiority over established baseline techniques.

While these initial results are encouraging, we recognize several avenues for future refinement. The performance of our method is inherently linked to the quality of the constructed similarity graph, which in this study relied on pre-trained Gecko embeddings [17] and a thresholding strategy. We believe that further exploration and optimization of graph construction techniques, including alternative embedding spaces and graph building algorithms, hold the potential for significant performance gains. Furthermore, the generation of representative synthetic data via prompt engineering is a crucial aspect, and while our current implementation demonstrates effectiveness, we anticipate that more sophisticated prompting strategies could yield even higher-quality synthetic examples, thereby further enhancing the overall performance of SYNAPSE-G. Finally, to solidify the generalizability of our findings, future work will involve a more extensive evaluation across a diverse range of datasets and rare event types. These directions, including investigations into more scalable graph learning approaches, represent promising next steps in advancing the capabilities of synthetic data-augmented semi-supervised learning for rare event classification.

References

[1] Shumeet Baluja, Rohan Seth, Dharshi Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. 2008. Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the 17th international conference on World Wide Web*. 895–904.

[2] James Banks. 2010. Regulating hate speech online. *International Review of Law, Computers & Technology* 24, 3 (2010), 233–239.

[3] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. 2006. Label Propagation and Quadratic Criterion. *Semi-Supervised Learning* (2006), 193–216.

[4] Camilla Casula, Sebastiano Vecellio Salto, Alan Ramponi, Sara Tonelli, et al. 2024. Delving into Qualitative Implications of Synthetic Data for Hate Speech Detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 19709–19726.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a Good Data Annotator?. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

[7] Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. 2024. Better Synthetic Data by Retrieving and Transforming Existing Datasets. *arXiv preprint arXiv:2404.14361* (2024).

[8] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496* (2022).

[9] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120.

[10] Matthew Herland, Richard A Bauder, and Taghi M Khoshgoftaar. 2019. The effects of class rarity on the evaluation of supervised healthcare fraud detection models. *Journal of Big Data* 6 (2019), 1–33.

[11] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 113–122.

[12] Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human Feedback is not Gold Standard. In *The Twelfth International Conference on Learning Representations*.

[13] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118* (2021).

[14] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

[15] Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277* (2020).

[16] Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. 2023. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684* (2023).

[17] Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. 2024. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327* (2024).

[18] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300* (2019).

[19] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*.

[20] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*. 173–182.

[21] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191* (2020).

[22] Sujith Ravi and Qiming Diao. 2016. Large scale distributed semi-supervised learning using streaming approximation. In *Artificial intelligence and statistics*. PMLR, 519–528.

[23] Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*. Springer, 232–241.

[24] Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*. 83–94.

[25] Chathurangi Shyalika, Ruwan Wickramarachchi, and Amit P Sheth. 2024. A comprehensive survey on rare event prediction. *Comput. Surveys* 57, 3 (2024), 1–39.

[26] Alexandra A Siegel. 2020. Online hate speech. *Social media and democracy: The state of the field, prospects for reform* (2020), 56–88.

[27] Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, et al. 2024. Beyond Human Data: Scaling Self-Training for Problem-Solving with Language Models. *Transactions on Machine Learning Research* (2024).

[28] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic

compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.

[29] Victor Suarez-Lledo and Javier Alvarez-Galvez. 2021. Prevalence of health misinformation on social media: systematic review. *Journal of medical Internet research* 23, 1 (2021), e17187.

[30] Amarnag Subramanya, Slav Petrov, and Fernando Pereira. 2010. Efficient graph-based semi-supervised learning of structured tagging models. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. 167–176.

[31] Partha Talukdar and William Cohen. 2014. Scaling graph-based semi supervised learning to large number of labels using count-min sketch. In *Artificial Intelligence and Statistics*. PMLR, 940–947.

[32] Sasan Tavakkol, Max Springer, Mohammadhossein Bateni, Neslihan Bulut, Vincent Cohen-Addad, and MohammadTaghi Hajiaghayi. 2025. Less is More: Adaptive Coverage for Synthetic Training Data. *arXiv preprint arXiv:2504.14508* (2025).

[33] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808* (2020).

[34] Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. *ProQuest number: information to all users* (2002).

A Omitted Proofs

A.1 Proof of Proposition 1

We first reprint the proposition for readability.

Proposition 2. *Let $Q \coloneqq S \cup N(S_+)$ denote the queried vertices and P be the number of positive examples in Q. Then,*

$$\mathbb{E}\left[\frac{P}{|Q|} \mid S\right] = (2q_1 - q_2) + \frac{1 + q_2\left(d + \frac{1}{p}\right) - q_1\left(d + \frac{2}{p}\right)}{\frac{1-p}{p} + h(S_+)}$$

$$\mathbb{E}\left[\frac{P}{|V|} \mid S\right] = \frac{p|S|}{|V|}\left((1 - 2q_1 + q_2) + (q_2 - q_1)d + (2q_1 - q_2)h(S_+)\right)$$

where $\frac{P}{|Q|}$ is the precision and $\frac{P}{|V|}$ is the recall.

Proof. $Q = S \cup N(S_+) = S_+ \cup S_- \cup N(S_+) = S_- \cup N(S_+)$. Since S is an independent set (Assumption 1), $S_- \cup N(S_+)$ is disjoint. Thus,

$$\begin{aligned} |Q| &= |S_-| + |N(S_+)| \\ &= (1-p)|S| + h(S_+)p|S| \\ &= (1 - p + ph(S_+))|S|. \end{aligned}$$

Let $S_1 \subseteq N(S_+) \setminus S_+$ be vertices in $N(S_+) \setminus S_+$ adjacent to exactly one vertex in S_+, and S_2 be those adjacent to exactly two. Let P_1, P_2 be the number of positive examples in S_1, S_2, respectively.

$$\begin{aligned} \mathbb{E}[P \mid S] &= \mathbb{E}[|S_+| + P_1 + P_2 \mid S] \\ &= |S_+| + q_1|S_1| + q_2|S_2|. \end{aligned}$$

We have:

$$|S_+| + |S_1| + |S_2| = |N(S_+)| \tag{1}$$

$$d|S_+| + |S_+| - |S_2| = |N(S_+)| \tag{2}$$

$$|N(S_+)| = |S_+|h(S_+). \tag{3}$$

Equation (1) counts vertices in $N(S_+)$. Equation (2) counts edges between S_+ and $N(S_+)$, subtracting $|S_2|$ once (as each is counted twice). Equation (3) is from the definition of $h(S_+)$. Solving (1)-(3):

$$\begin{aligned} |S_1| &= (2h(S_+) - d - 2)|S_+| \\ |S_2| &= (d + 1 - h(S_+))|S_+| \\ \mathbb{E}[P \mid S] &= |S_+|(1 + q_1(2h(S_+) - d - 2) \\ &\quad + q_2(d + 1 - h(S_+))) \\ &= p|S|((1 - 2q_1 + q_2) \\ &\quad + (q_2 - q_1)d + (2q_1 - q_2)h(S_+)). \end{aligned}$$

Therefore,

$$\begin{aligned} &\mathbb{E}\left[\frac{P}{|Q|} \mid S\right] \\ &= \frac{p|S|((1 - 2q_1 + q_2) + (q_2 - q_1)d + (2q_1 - q_2)h(S_+))}{(1 - p + ph(S_+))|S|} \\ &= (2q_1 - q_2) + \frac{1 + q_2\left(d + \frac{1}{p}\right) - q_1\left(d + \frac{2}{p}\right)}{\frac{1-p}{p} + h(S_+)}. \end{aligned}$$

If $1 + q_2\left(d + \frac{1}{p}\right) - q_1\left(d + \frac{2}{p}\right) > 0$, then we must have that

$$p > \frac{2q_1 - q_2}{1 + (q_2 - q_1)d}$$

and $\mathbb{E}\left[\frac{P}{|Q|} \mid S\right]$ decreases with $h(S_+)$. Now, we compute the derivative of $\mathbb{E}\left[\frac{P}{|Q|} \mid S\right]$ with respect to p:

$$\begin{aligned} &\frac{\partial}{\partial p}\mathbb{E}\left[\frac{P}{|Q|} \mid S\right] \\ &= \frac{(1 - q_1(2 + d) + q_2(1 + d)) + h(S_+)(2q_1 - q_2)}{(1 - p + ph(S_+))^2} \\ &= \frac{1 + d(q_2 - q_1) + (h(S_+) - 1)(2q_1 - q_2)}{(1 - p + ph(S_+))^2}. \end{aligned}$$

Since $h(S_+) \le d + 1$, $d \ge h(S_+) - 1$. Thus,

$$\begin{aligned} &\frac{\partial}{\partial p}\mathbb{E}\left[\frac{P}{|Q|} \mid S\right] \\ &\ge \frac{1 + (h(S_+) - 1)(q_2 - q_1) + (h(S_+) - 1)(2q_1 - q_2)}{(1 - p + ph(S_+))^2} \\ &= \frac{1 + (h(S_+) - 1)q_1}{(1 - p + ph(S_+))^2} > 0. \end{aligned}$$

Therefore, $\mathbb{E}\left[\frac{P}{|Q|} \mid S\right]$ is strictly increasing with respect to p.

Finally,

$$\begin{aligned} \mathbb{E}\left[\frac{P}{|V|} \mid S\right] = \frac{p|S|}{|V|}\big((1 - 2q_1 + q_2) \\ + (q_2 - q_1)d + (2q_1 - q_2)h(S_+)\big). \end{aligned}$$

Since $q_2 < 2q_1$, $\mathbb{E}\left[\frac{P}{|V|} \mid S\right]$ increases with p and $h(S_+)$. □

Received 03 June 2025