# SMIDGen: An Approach for Scalable, Mixed-Initiative Dataset Generation from Online Social Networks

**Matthew Louis Mauriello[1], Cody Buntain[2], Brenna McNally[2]**
**Sapna Bagalkotkar[1], Samuel Kushnir[1], Jon E. Froehlich[1]**
Makeability Lab | Human-Computer Interaction Lab (HCIL)
Department of Computer Science[1], College of Information Studies[2]
University of Maryland, College Park
{mattm401, cbuntain, bmcnally}@umd.edu

## ABSTRACT

Recent qualitative studies have begun using large amounts of Online Social Network (OSN) data to study how users interact with technologies. However, current approaches to dataset generation are manual, time-consuming, and can be difficult to reproduce. To address these issues, we introduce SMIDGen: a hybrid manual + computational approach for enhancing the replicability and scalability of data collection from OSNs to support qualitative research. We demonstrate how the SMIDGen approach integrates information retrieval (IR) and machine learning (ML) with human expertise through a case study focused on the collection of YouTube videos. Our findings show how SMIDGen surfaces data that manual searches might otherwise miss, increases the overall proportion of relevant data collected, and is robust against IR/ML algorithm selection.

## Author Keywords

Qualitative data collection; mixed-initiative; social media; user-generated content; machine learning; query expansion

## ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

## INTRODUCTION

Online Social Networks (OSNs) such as *Twitter*, *YouTube*, and *reddit* have emerged as valuable data sources for qualitative studies of everyday interactions with technology [2,3,6,14]. By studying user-generated content, researchers get access to naturalistic data about end-users and populations that are otherwise challenging to observe [16]. However, modern OSNs generate millions of content pieces and hundreds of hours of video every minute [22]. Researchers face challenges related to scale, noise filtering [20], rapidly evolving vocabularies that hinder comprehensive searches [11], and restricted access to proprietary platforms (*e.g.*, rate limits on queries) [10,19]. Typically, these challenges are addressed through time-intensive manual searches, often costing hundreds of researcher-hours [2,6], or focusing on small, downsampled datasets (*e.g.*, 100 videos [3]) that risk missing insights or misrepresenting a domain or topic.

To assist researchers in constructing OSN-based datasets for large-scale qualitative analysis, we introduce *SMIDGen: A Scalable, Mixed-Initiative Dataset Generation* approach. SMIDGen combines algorithms in information retrieval (IR) and machine learning (ML) along with a traditional qualitative coding process to assist with data collection and filtering. SMIDGen has four phases: (i) manually exploring an OSN and generating keywords to bootstrap data collection, (ii) computationally expanding these queries to increase domain/topic coverage, (iii) mixed-initiative data labeling and training to construct automated models, and (iv) applying these models at scale to generate a final dataset that is larger and more diverse as a result.

After describing each of these phases, we demonstrate their application and utility through a detailed use case on YouTube: studying non-professional "everyday uses" of thermal cameras. Our findings suggest that the automated query expansion in Phase 2 contributes new data that we would have otherwise missed, and the classification models from Phases 3 and 4 accurately identified domain and topic relevance. We also show that the SMIDGen approach is robust against algorithm selection, which facilitates implementation, and that one need not manually label an entire dataset to achieve performance enhancements. We close with a discussion of SMIDGen and OSN data collection highlighting key strengths, limitations, and suggestions for improving performance.

## QUALITATIVE STUDIES OF OSN CONTENT

Research involving data from OSNs generally derives insights from quantitative analyses of word frequencies, network structures, and other measurable artifacts [9,13]. However, recent studies have demonstrated the value of harnessing user-generated content (*e.g.*, videos, images) as a source of naturalistic data for large-scale qualitative research on how end-users interact with technologies [2,3,6,14]. The topic areas and networks approached in these studies are diverse, ranging from studying assistive technologies on *YouTube* [2] and *Thingiverse* [6] to political discourse on *Twitter* [7]. While differences exist in how data is queried and collected, each apply a similar high-level method: (i) a researcher explores an OSN, becoming familiar with the target domain, and how it is discussed, to generate an initial set of keywords; (ii) these keywords are used as search terms—either individually or in combination—on the OSN; (iii) researchers manually compile relevant artifacts from the first few hundred search results, potentially extracting new keywords; (iv) steps *(ii)* and *(iii)* repeat until enough artifacts are collected, search terms are exhausted, or saturation is reached. The resulting documents comprise the final dataset (Figure 1, top).

These approaches have a high cost in researcher hours and are therefore difficult to scale. Furthermore, their reliance on proprietary browser-based search interfaces and their platform-specific nature (*i.e.,* being derived for Twitter) make them difficult to reproduce or apply to other OSNs. While many tools such as NVivo™ and Atlas.ti™ exist to support the *analysis* of collected data, resources for data collection—including tools and guidance on approaches [12]—are far scarcer. To address this gap, in this work we outline an approach that applies hybrid manual + computational techniques to assist with data collection and relevance filtering.

## AN OVERVIEW OF THE SMIDGEN METHOD

SMIDGen is a mixed-initiative dataset generation approach that combines algorithms in IR and ML along with traditional qualitative coding processes to semi-automatically expand study datasets, ensure high relevance, and reduce manual labeling efforts. While SMIDGen is intended to be OSN and research domain agnostic, the method makes two assumptions: (i) the researcher has a specific, observable domain or topic of interest and (ii) the researcher is using a specific OSN with a query-able API. Below, we provide a high-level description of the four-phase approach before providing a specific use case that illustrates SMIDGen's utility in practice.

### Phase 1: Data Exploration and Initial Keyword Creation

Phase 1 begins similarly to the large-scale qualitative studies found in recent literature [2,6,14]. The research team performs an informal investigation of a target OSN to gain familiarity with the platform and research target. The goal is, first, to understand the platform-specific features (*e.g.,* hashtags) and restrictions (*e.g.,* rate limits) of the target OSN. The second goal is to identify search keywords based on relevant web domains, acronyms, phrases, or hashtags that appear often within the target domain area and can be used to gather relevant data. These initials keywords should emphasize breadth—covering as much information pertaining to the domain of interest as possible. The researcher should then construct a small, preliminary dataset of relevant documents.

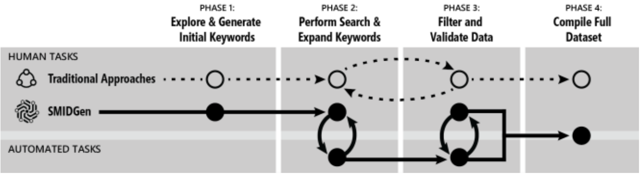


**Figure 1.** A diagrammatic overview of the SMIDGen method compared to traditional approaches to support large-scale analyses of OSN-based datasets. SMIDGen combines manual search with computational methods to semi-automatically expand research datasets and ensure high relevance.

### Phase 2: Computational Query and Dataset Expansion

The varied and evolving vocabularies in OSNs [8] and the constrained (or rate-limited) access to their raw data suggests this preliminary dataset is likely incomplete. The approach used in prior work had researchers manually analyze this preliminary data to construct an exhaustive set of all (known) relevant keywords. SMIDGen's Phase 2 accelerates and expands this step by applying IR algorithms on the initial data from Phase 1 to generate new, relevant, unanticipated search queries automatically. This query expansion process may leverage thesauruses to identify synonymous keywords [9], data-driven approaches like word embeddings [18], or relevance feedback that identify common keywords in relevant search results [5]. In choosing an expansion technique, researchers must consider whether they have unrestricted access to the underlying OSN data (rarely the case in OSNs) and search term semantics (*e.g.,* thesauruses are less useful for finding synonyms for proper nouns). Common approaches are based on co-occurring frequencies or statistical language models—both of which we evaluate in our use case. Once the query expansion algorithm is selected and executed, each term from the expanded set of keywords is queried to generate a larger, more exhaustive dataset. Since this phase prioritizes comprehensiveness this intermediate dataset is likely noisy, which is addressed next.

### Phase 3: Mixed-Initiative Analysis and Modeling

Depending on the OSN and domain of interest, Phase 2's resulting dataset may contain thousands to millions of results, some of which are likely irrelevant. In prior work datasets were filtered manually, which is time intensive. For example, in [6], Buehler *et al.* spent hundreds of researcher-hours generating a dataset of only a few thousand artifacts. SMIDGen accelerates this coding in Phase 3 via a mixed-initiative process in which researchers code small, tractable samples (*e.g.,* a few hundred) to train ML models, which then apply these codes automatically. Researchers then manually validate samples of these machine labels (by applying their own labels to the samples and comparing) to ensure the model's output is reasonable. If human-machine agreement is satisfactory then the machine labeled data is passed to the final phase of SMIDGen. If human-machine agreement is below a researcher-established threshold, the new human labels are fed back into the models for retraining and researchers manually validate new samples of data. Researchers should perform this feedback loop over until they are confident in the ML models.

### Phase 4: Classifier Application and Dataset Assembly

Following Phase 3, the researcher is left with a selection of manually analyzed artifacts, manually validated ML models, and a dataset of artifacts with machine labels. In this final stage, the researcher can save these models for initializing future data collection tasks and constructs the final, relevant dataset for subsequent qualitative analysis by combining the human-labeled and machine-inferred data.

## APPLYING SMIDGEN IN PRACTICE

To more deeply illustrate SMIDGen, we offer a specific use case: searching for and qualitatively analyzing an emergent technology on YouTube—specifically, non-professional "everyday" uses of thermal cameras (see [*In Review*]).

### Phase 1: Data Exploration and Initial Keyword Creation

Recall that the goals of Phase 1 are to familiarize oneself with the OSN and the domain of interest, generate a list of initial keywords, and query these keywords to construct an initial dataset. To begin, we queried the quoted string "thermal camera" on YouTube's website both alone and in combination with other common thermography-related terms (*e.g.,* "surveillance", "medical"). We manually assessed these search results to identify relevant videos and generate an initial list of seven keywords (Table 1, top row). Using a custom Python script, we then queried these terms via the YouTube Data API (v3) to create a preliminary study dataset. Following recommendations of Anthony *et al.* [2], we extracted the first 200 YouTube results for each term and stored the resulting video URL and metadata (title, description, author, view counts, *etc.*). In total, we collected 1,400 videos, which was reduced to 1,092 after removing duplicates. This preliminary dataset provided input to the query expansion algorithms in Phase 2 and constituted a subset of the final dataset in Phase 4.

### Phase 2: Computational Query and Dataset Expansion

The goal of Phase 2 is to automatically increase the set of relevant data beyond what could easily be found with manual search. To do this, Phase 2 leveraged the initial dataset from Phase 1 to automatically identify additional query terms—a process known as query expansion [9]. In our case, we used video titles and descriptions from our Phase 1 dataset as input to our query expansion algorithms. Typically, a researcher would employ only a single expansion technique; however, as an early exploration of SMIDGen we compared three different approaches (see Evaluation). Applying all three approaches resulted in a fourfold increase in dataset size (> 4,000 videos); however, expanded datasets are typically noisy—a known side effect of query expansion—necessitating a scalable method to remove irrelevant data.

### Phase 3: Mixed-Initiative Analysis and Modeling

Phase 3 provided computational methods to support video classification in the Phase 2 dataset, which contained videos unrelated to thermography and thermographic videos not relevant to our specific research interest (*e.g.,* marketing). This classification required an initial labeled dataset for classifier training so that the classifier could then apply those labels to the remaining data automatically. To create our training set, we qualitatively coded the Phase 1 dataset using a traditional iterative coding process [5,15]. Informed by Blythe and Cairn's study [3] of iPhone use on YouTube, we generated an initial codebook with 7 codes. Two research assistants independently coded 200 randomly selected Phase 1 videos using titles, descriptions, and video content as input. Each video with a single category, and we used Cohen's kappa to calculate inter-rater reliability (IRR). After three rounds of coding, IRR was 0.69, rated *good agreement* [24]. The remaining Phase 1 data ($N$=492) was then divided equally and coded individually, resulting in codes for all 1,092 Phase 1 videos.

To utilize this labeled dataset for training an ML classifier, we featurized the textual and authorship aspects of each YouTube video in our dataset. Text featurization converted video titles and descriptions into a bag-of-words model and weighting terms using term-frequency, inverse-document-frequency (TF-IDF). Authorship featurization converted a video's author into a categorical feature—an ML ensemble process called stacking [23]—which we included to capture information like whether a video was posted by a company's marketing account. While the image and audio data in the videos themselves could also be used an input vectors, we did not explore this in our work. For the classifier itself, we experimented with three approaches using Python's Scikit-Learn library [21]: logistic regression, random forests, and a support vector machine (see Evaluation). We also modeled video classification as two binary tasks: relevant-versus-non-relevant and "everyday use"-versus-other.

Regardless of classifier type, one would expect a classifier trained on Phase 1 data to classify some videos in the Phase 2 dataset incorrectly given the new features introduced by query expansion (*e.g.,* new authors or words in titles and descriptions). To account for this possibility, after applying the Phase 1-trained classifier to Phase 2 data, we manually labeled a subset of data in the Phase 2 dataset—which served as ground truth. Our two research assistants separately annotated four batches of 250 videos in this Phase 2 sample, using the same process as with the Phase 1 data. After each batch, we validated our classifiers against this labeled data and evaluated feature performance. In total, we randomly sampled and qualitatively coded 2,090[1] videos from the Phase 2 dataset; however, this large amount would ordinarily be unnecessary and was chosen to study the effect of training set size on classifier performance. Finally, we trained our classifiers on all of our labeled data and performed a final spot check on results to validate our metrics: We randomly sampled videos from the set of

[1] Some videos were separated from those labeled by our (high school) research assistants and labeled by other (adult) researchers because they potentially contained graphic content based on keywords. They were included in training to maintain topic diversity in the training dataset.

automatically labeled data, 100 videos each from both classes and binary classification tasks, for a total of 400 videos. Researchers manually labeled 200 videos from each task without knowing the classifier's inferred labels, which we checked then checked against classifier metrics.

### Phase 4: Classifier Application and Dataset Assembly

Finally, in Phase 4 we applied the validated ML classifier to the remaining unlabeled videos in our Phase 2 dataset. Our final dataset of thermographic relevant videos (N=4,380) included both human-labeled (N=2,082) and machine-labeled (N=2,298 videos) data combined from Phases 1-3, and our final dataset of "everyday use" videos (N=1,686) also included human-labeled (N=772) and machine-labeled (N=914) data. This "everyday use" dataset was then used to study end-user experiences with everyday uses of commodity thermographic technology (see [*In Review*]).

## EVALUATING SMIDGEN

To ensure the SMIDGen approach expanded datasets and accelerated researchers' domain relevance and subtopic classification, we conducted an extended evaluation using the above use case. While future researchers may not need to conduct such thorough investigations themselves, this validation was critical in establishing confidence in the SMIDGen approach and conducting similar evaluations would is an important error-analysis step for understanding systematic issues and when collecting data on different topics or from different platforms. These assessments relied on the *precision* and *recall* metrics common in IR and ML. In this context, precision was defined as the ratio of retrieved, relevant videos to all retrieved videos. Recall would typically be the ratio of all retrieved, relevant videos to all possible relevant videos that exist on YouTube; however, we used a modified recall metric described below. Specifically, we evaluated the following research questions:

**RQ1.** How well did query expansion work to expand the initial dataset and find additional relevant videos?

**RQ2.** How accurately did Phase 3's classification models identify domain-relevant videos (*i.e.,* videos about thermal cameras)?

**RQ3.** Within the domain relevant dataset, how accurately did classification models identify the subtopic of "everyday use"?

**RQ4.** How much manual labeling was required for accurate automatic domain relevance and topic identification?

### RQ1: How Well Did Query Expansion Work?

The perfect query expansion algorithm would capture all remaining relevant data on the target OSN (recall) and all data returned would be relevant (precision). For most OSNs, however, calculating recall is difficult or impossible given restricted OSN database access and large data volumes, where manually labeling millions of documents is infeasible [4]. We therefore introduce a new metric, called an *expansion coefficient* (Eq. 1), where $t_e$ is the number of relevant items that exist in the expanded dataset but *not* in the initial dataset, and $t_o$ is the intersection of relevant items in both datasets. This metric ranges from [0,1], with a perfect expansion score of 1 indicating every relevant item captured by the expansion would **not** have been captured in the original dataset.

| | Search Terms | Videos |
|---|---|---|
| **Seed** | infrared, lepton, thermal, thermal camera, thermal image, thermal imaging, thermography | 1,092 |
| **COO** | flir lepton, flir one, flir thermal, follow u, imaging camera, infrared camera, infrared thermography, night vision, thermal imager, u facebook | 4,264 |
| **KLD** | breast thermography, flir lepton, flir one, flir thermal, imaging camera, infrared thermography, night vision, seek thermal, thermal imager, thermal paste | 5,075 |
| **HITL** | breast thermography, flir lepton, flir one, flir thermal, imaging camera, infrared camera, infrared thermography, night vision, seek thermal, thermal imager | 4,670 |

**Table 1.** The initial search term list from Phase 1, which was manually created (Seed) along with the expanded keyword set for each query expansion technique. The Videos column represents the total number of videos returned by querying the search terms on YouTube; while duplicates were removed within each set, duplicates exist across sets.

$$\textit{Expansion Coefficient} = t_e \,/\, (\, t_e + t_o \,) \qquad \text{(Eq. 1)}$$

As multiple standard query expansion approaches exist in IR, we tested three: a frequency-based algorithm using word co-occurrence (COO), a statistical language model using Kullback-Leibler divergence (KLD) [17], and a human-in-the-loop method (HITL) in which a researcher manually selected search terms from a list of suggestions from COO and KLD. Moving from COO to KLD to HITL represented an increase in complexity, with COO requiring only the Phase 1 dataset, KLD requiring additional data on YouTube's general language, and HITL requiring human expertise to select the most salient search terms.

As input, both COO and KLD used the Phase 1 video titles and descriptions (*N*=1,092 videos) as well as the original seven search terms. As output, COO and KLD generated rankings of two-word phrases by scoring their frequencies (the more common phrases were ranked higher). While COO used this raw frequency as its score, KLD modulated scores for frequent-but-generic phrases common in a random sample of 51,889 YouTube videos, acquired from YouTube's 8m dataset [1]. We then removed keywords already present in the initial Phase 1 keyword set and used the remaining top-ten[2] ranked two-word phrases as expansion terms. For the HITL method, we presented the union of these top terms to the research team, who then selected terms they thought to be the most salient. The top ten suggested search terms for each approach are shown in Table 1. Notably, COO and KLD had high overlap in their suggestions with only 13 search terms being unique.

To build the expanded dataset, we constructed a query set composed of the expanded search terms and all pairwise combinations with Phase 1's initial terms (*e.g.,* we searched for "thermal imager" and "infrared AND 'thermal imager'"), as per Anthony *et al.* [2], and queried YouTube's API for each new query. Similar to [2], we stored the first 200 videos for each query. After removing intra-set

[2] Selected to minimize researcher effort in this phase.

| | Seed | COO | KLD | HITL |
|---|---|---|---|---|
| **# of Manually Labeled Videos** | 1,092 | 2,010 | 2,352 | 2,164 |
| **Precision** | 0.59 | 0.81 | 0.79 | **0.88** |
| **Expansion Coefficient** | — | 0.72 | 0.73 | **0.74** |

**Table 2.** The amount of manually labeled data used in our experiments and the results from our query expansion evaluation with our proxy recall metric.

| | Manually Labeled | | Automatically Labeled |
|---|---|---|---|
| | **Seed** | **Expanded** | **Expanded** |
| **Relevant** | 647 | 1913 | 4134 |
| **Total** | 1092 | 2164 | 4626 |
| **Ratio** | 0.5925 | 0.884 | 0.8936 |

**Table 3.** Relevant videos using both manual and automated labeling methods. Ratios of relevant videos are consistent across manual and automated methods.

| | Relevant Accuracy | Irrelevant Accuracy | Everyday-Use Accuracy | Other-Use Accuracy |
|---|---|---|---|---|
| **Rater 1** | 79/100 | 97/100 | 70/100 | 88/100 |
| **Rater 2** | 92/100 | 93/100 | 70/100 | 89/100 |
| **Resolved** | 94/100 | 91/100 | 70/100 | 89/100 |

**TABLE 4.** Results from final manual classifier validation. Accuracy estimates predicted during classifier training were consistent with accuracy in manual validation for both domain- and topic-relevance.

duplicates, COO, KLD, and HITL expanded queries resulted in 4,264, 5,075, and 4,670 videos respectively (Table 1), though all expanded sets had many videos in common. The two key questions remain: how many *new* videos were found not in the original dataset (as measured by the expansion coefficient) and how many were *relevant*.

To address both questions, we used the 1,092 labeled videos from Phase 1's initial dataset and manually labeled a random subset of Phase 2's expanded dataset. Given the large query term intersection among the three expansion methods, intersection among search result was consequently large. Since the likelihood of selecting duplicate videos across the expansion datasets was therefore high, we pooled these expansion results together and randomly selected and labeled videos from this dataset (2,090 videos exclusively from the expanded set, 3,182 videos in total). The results are shown in Table 2. Notably, all three approaches achieved similar expansion scores—between 0.72 and 0.74—with HITL performing best. That is, approximately three of every four relevant videos captured by our expansion approach were new and would otherwise have been missed in the initial dataset. But were these new videos relevant? If not, then the query expansion process will increase researcher burden as they filter through noisy data. Fortunately, in terms of precision, again all three approaches performed well—increasing precision by over 34% compared to Phase 1's initial dataset with HITL again performing best at 0.88.

The key implication here is that, regardless of algorithm used, query expansion substantially increased dataset sizes for both raw count and relevant videos. Volume of relevant videos more than doubled, with expansion coefficients showing the majority of relevant videos captured in each expansion was omitted from the initial dataset. The HITL query expansion process further increased these sizes as well, demonstrating the value of integrating human selection into the expansion process.

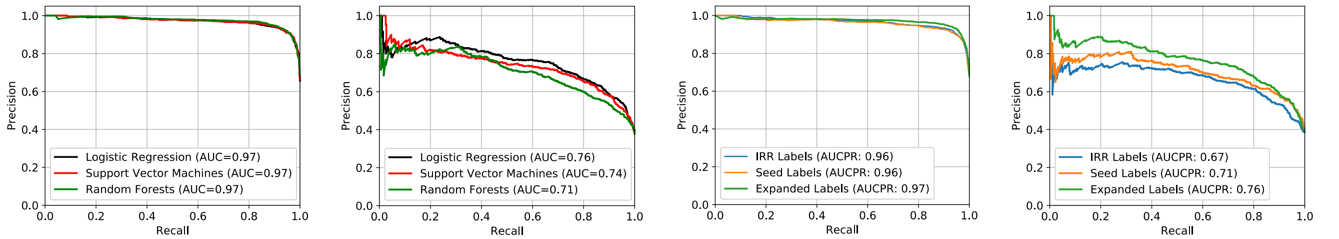
### RQ2: How Well Did Phase 3 Classify Domain-Relevant Videos?

After query expansion, we are left with several thousand unlabeled videos. To determine whether ML algorithms could accurately classify these videos as relevant (thermographic) or irrelevant (non-thermographic), we evaluated three text-based classification algorithms: logistic regression (LR) from probabilistic models, random forests (RF) from tree-based models, and support vector machines (SVMs) for geometric models. By choosing one model from each of the main families of classifiers, we could further examine the effect of classifier type on performance.

We trained each classifier on titles, descriptions, and video authors using all our manually labeled data, consisting of 3,182 videos (2,082 marked as relevant, 1,100 irrelevant). To evaluate each classifier, we used 10-fold cross-validation where a fold was a random 10% of the 3,182 labeled videos. Our primary measure was the area under the precision-recall curve (AUPRC), selected because of its robustness against class imbalance. While an ideal classifier would achieve an AUPRC of 1.0, we expected the three classifiers would perform well classifying domain relevance given prior work [8].

Our results are shown in Figure 2a. We found that each ML model exhibited high performance, with RFs performing marginally better than other models. Feature analysis compared text-only classifiers (*i.e.,* using title and description) to author-only and text-and-author classifiers, with textual features alone outperforming author-based classification by 10% and performing marginally better than text-and-author versions. Regardless of the ML model—SVMs, LR, or RFs—all classifiers performed well, achieving AUPRC scores above 0.97 and a mean accuracy of 91.3%. These high AUPRC scores suggest the decision boundary between relevant and irrelevant videos was not complex, which one might expect given sufficiently clear queries. Table 3 shows proportions of both manually and automatically labels videos, demonstrating consistency between proportions of manual and automated labels.

To ensure that our models' AUPRC metrics aligned with researcher assessments, we performed a final external validation step by having two researchers manually inspect a random sample of the automatically labeled data. Ideally, these researchers' labels would agree with the inferred labels. Since the RF classifier was the highest performing domain-relevance classifier, we applied it to the remaining unlabeled videos, resulting in automatic labels of 808 irrelevant videos and 2,298 as relevant to thermographic use. As described in the Phase 3 use case above, we randomly sampled 100 videos from the set of classified-relevant videos and 100 from the classified-irrelevant videos. Two research assistants manually labeled these videos and resolved conflicts, as shown in Table 4. We



**(a)** Effect of classifier on domain classification performance

**(b)** Effect of classifier on subtopic classification performance

**(c)** Effect of label set size on performance of RF classifier for domain classification

**(d)** Effect of label set size on performance of LR classifier for subtopic classification

**Figure 2.** Results for RQ2 (a), RQ3 (b), and RQ4 (c and d). In sum, the RF classifier performed best in domain classification, while the LR classifier performed best at subtopical; label size had negligible impact on domain classification, but improved subtopical classification.

found this classifier achieved 92.5% accuracy, consistent with but slightly higher than our accuracy estimates.

There are two key implications here: first, for automatically identifying relevant videos, SMIDGen was robust against classification algorithm selection, performing well despite the family of algorithm used, with textual features being sufficient for classifying relevance. Second, automatically classified labels were consistent with manual labels in both distribution and manual validation, so cross-validation and optimizing AUPRC were valid training strategies here.

### RQ3: How Well Did Phase 3 Automate Subtopic Classification?

Having established models for automatically inferring relevance, we moved to determining whether similar ML models could identify specific subtopical content *within* the set of relevant videos. To investigate this, we focused on classifying videos as "everyday use of thermal cameras" or not (a binary classification). Using the manually labeled dataset from Phase 2 of 2,298 relevant videos (772 everyday-use videos, 1,130 non-everyday-use thermographic videos) as ground truth data, we repeated the above ML pipeline. In comparison to domain classification, we expected this task to be more difficult as the decision boundary between subtopics is typically more complex.

Given that "everyday use" videos were underrepresented in this training data, we weighted "everyday use" videos at a rate of 3:1 (an estimate of the expected underrepresentation) during training, such that a false negative in the "everyday use" class was penalized three times more than a false positive. After training and again using 10-fold cross-validation, the average model score was 0.73 AUPRC (Figure 2b), with LR performing the best (0.76 AUCPR and 79.39% accuracy). Unlike relevance prediction, we also found that including video authorship features increased performance. We then applied the LR model to the 2,298 unlabeled-but-classified-as thermographic videos, which classified 902 of these videos as "everyday use."

As in the relevance evaluation, we manually validated this topical model by randomly sampling an additional 100 "everyday use" videos and 100 "non-everyday use" videos (*e.g.*, professional marketing videos). Table 4 shows researcher assistant-labeling results, which were consistent with automatic classification (79.5% *vs.* predicted 79.4%) and our expectations of a more difficult ML task. Furthermore, the higher accuracy in non-everyday-use is consistent with our class weighting, which would prefer these type-II errors over type-I errors. The main result here is that subtopical classification is more difficult but still feasible. We also note algorithm and feature engineering had higher impact on model performance in this context.

### RQ4: How Much Manual Labeling Effort is Necessary?

Finally, while the above tasks used approximately 3,000 videos for training, we wanted to investigate how much manual annotation was necessary to achieve the above results. We therefore performed two experiments on training data size: one on classifying domain-relevance and one on classifying subtopics. An expectation here might be that fewer manual labels would be necessary for classifying relevance, but more everyday-use labels would increase subtopic classifier performance. For both experiments, we retrained the highest-performing classifier configurations from RQ2 and RQ3 on three datasets of increasing size: a subset of Phase 1's initial seed dataset that we used for evaluating IRR (*N*=571), the full Phase 1 dataset (*N*=1,092), and the full labeled data from Phase 1 and Phase 2 (*N*=3,182). In the ideal case, models trained on the 571 data points would perform as well as the models trained on the full dataset substantially reducing researcher effort.

Figure 2c shows these results for the RF-text-only classifier for domain relevance, in which increasing dataset size had little impact. Figure 2d shows results for the LR-text-and-author classifier for subtopics, and we see here that topic identification (*i.e.*, "everyday use") did benefit from additional training. Moving from the IRR (*N*=340) dataset to the Phase 1 dataset (*N*=647) and from the Phase 1 dataset to the Phase 2 dataset (*N*=2,298) resulted in ~5% and ~10% increases in performance, respectively.

The key result here is that, researchers can achieve high performance in automated relevance filtering using a small dataset that they would likely already have developed for IRR (using traditional qualitative coding methods). For subtopics classification, however, more manual effort translates to better classifiers. This additional effort would be offset by the reduction in domain-irrelevant videos.

## DISCUSSION, LIMITATIONS, AND FUTURE WORK

In this paper, we presented SMIDGen, a hybrid manual + computational approach for generating OSN-based data for qualitative research. Where possible, SMIDGen leverages research steps that would be performed in a more traditional qualitative study (*e.g.*, keyword generation, codebook validation, inter-rater reliability), and much of the remaining overhead can be automated to minimize researcher effort. An additional key benefit of SMIDGen is its replicability. Not only does SMIDGen allow for the collection of single snapshots of data available on OSNs at a given time, but artifacts from this process (*e.g.*, machine learning models and labeled data) can then be stored for future use to iterate upon or bootstrap future data collection. For future collection tasks, the researcher need only update the initial search queries to reflect new developments in the research domain. These new queries are integrated into Phase 2's query expansion and prior models are then retrained on updated labels from the more recent dataset.

One challenge to OSN-based research in general is that target platforms can limit insights and research strategies. In our use case, YouTube's API criteria to determine which videos are returned for a given query is unknown. It is therefore unclear whether relevance searches are biased towards recent videos, videos from users with many likes or subscribers, or some paid promotional scheme. The use of SMIDGen partially addresses this concern by running many queries with limited overlap to develop a more comprehensive dataset despite these unknown ranking schemes. These challenges are not specific to YouTube; most OSNs have similar result set constraints and opaque sampling methods [4]. While one could mitigate these issues with different query techniques, requirements are likely to change across platforms, so this dependence on platform capabilities is likely to continue unless one pays for partnerships with data providers.

Another challenge common to qualitative research, and which applies to SMIDGen, is dependence on human expert queries. It may be tempting to rely on query expansion to ensure comprehensive data collection, but a fundamental requirement is that query expansion assumes some overlap exists between the expert query and other related queries. If this overlap does not exist (*e.g.*, because of linguistic barriers or disjoint communities), SMIDGen may miss important subsets of data. As such, it is important for researchers to cover as many communities of discussion as possible with their initial queries (in Phase 1). These queries need not be complete, but should have good coverage of the domain as SMIDGen uses query expansion to enhance comprehensiveness but does not guarantee it.

As SMIDGen relies on ML models for scaling up analysis, an additional challenge is the way these models may bias results based on training data and feature selection. For example, in our use case, classifiers were trained with video title, description, and author, while human evaluators also assessed the video content itself. For identifying videos relevant to thermal camera use, this limitation seemed to have little effect but multimedia omission may explain the increased difficulty classifying the "everyday use" topic (RQ3). Such a limitation applies equally to other platforms as well: if one wanted to study Twitter or Instagram data for instance, textual models will ignore image data. One of SMIDGen's strengths, however, is the flexibility to extend the machine learning models. One can integrate additional features into the models as new features become available or new technology is developed (*e.g.,* computer vision or speech recognition approaches to analyze video/audio data).

Finally, this study presented a single, initial use case that leveraged only one OSN platform. Thus, future work should explore expansions to new domains/topics and platforms. While the query expansion and ML classifiers may be adequate for YouTube's API and our research focus (studying thermal camera use), more research is necessary to evaluate these methods and SMIDGen as a general approach particularly in light of recent advances (*e.g.*, [18]). To aid other researchers in applying the SMIDGen approach, we are actively working on developing a SMIDGen web application that should work across popular OSNs including YouTube, Instagram, and Twitter as well as performing additional evaluations that look at the tradeoffs in human/researcher effort and quality of results.

## CONCLUSION

In this paper, we have presented SMIDGen: a scalable, mixed-initiative approach for generating large-scale, comprehensive datasets for qualitative research. We provided both a high- and low-level description of this approach and evaluated it on a single use case on non-professional, everyday use of thermographic cameras. Our results provide guidance to researchers on applying SMIDGen to their own research including how to use: (i) query expansion methods to increase recall and reduce bias and (ii) ML models to assist with both relevance filtering and topic selection. However, this approach is preliminary, and its presentation here aims to showcase an interesting use case for IR and ML techniques in qualitative research. Additionally, throughout this work we have discussed key challenges associated with extracting data from OSNs and highlighted potential areas for improving the performance of this approach (through crowdsourcing, automated video analysis, *etc.*), which researchers and application designers may be interested in exploring further.

## ACKNOWLEDGEMENTS

While conducting this research, Sapna Bagalkotkar and Samuel Kushnir were interns in our group and attended Poolesville High School and Berman Hebrew Academy, respectively; both located in Maryland. Additionally, Cody Buntain received funding from the IC Postdoctoral Research Fellowship Program. Finally, Jon E. Froehlich recently joined the Paul Allen School of Computer Science & Engineering at the University of Washington.

## REFERENCES

1. Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.
2. Lisa Anthony, YooJin Kim, and Leah Findlater. 2013. Analyzing User-generated Youtube Videos to Understand Touchscreen Use by People with Motor Impairments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 1223–1232. https://doi.org/10.1145/2470654.2466158
3. Mark Blythe and Paul Cairns. 2009. Critical Methods and User Generated Content: The iPhone on YouTube. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09), 1467–1476. https://doi.org/10.1145/1518701.1518923
4. Praveen Bommannavar, Jimmy Lin, and Anand Rajaraman. 2016. Estimating Topical Volume in Social Media Streams. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing* (SAC '16), 1096–1101. https://doi.org/10.1145/2851613.2851810
5. V Braun and V Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2: 77–101.
6. Erin Buehler, Stacy Branham, Abdullah Ali, Jeremy J Chang, Megan Kelly Hofmann, Amy Hurst, and Shaun K Kane. 2015. Sharing is Caring: Assistive Technology Designs on Thingiverse. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 525–534. https://doi.org/10.1145/2702123.2702525
7. Cody Buntain and Jennifer Golbeck. 2017. I Want to Believe: Journalists and Crowdsourced Accuracy Assessments in Twitter. *arXiv preprint arXiv:1705.01613*.
8. Cody Buntain, Erin Mcgrath, and Brandon Behlendorf. 2018. Sampling Social Media: Supporting Information Retrieval from Microblog Data Resellers with Text, Network, and Spatial Analysis. In *51st Hawaii International Conference on System Sciences (HICSS)*.
9. D Manning Christopher, Raghavan Prabhakar, and SCHÜTZE Hinrich. 2008. Introduction to information retrieval. *An Introduction To Information Retrieval*.
10. Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. 2014. Assessing the bias in samples of large online networks. *Social Networks* 38: 16–27.
11. Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. 2011. Contextual Bearing on Linguistic Variation in Social Media. In *Proceedings of the Workshop on Languages in Social Media* (LSM '11).
12. Greg Guest, Emily E Namey, and Marilyn L Mitchell. 2012. *Collecting qualitative data: A field manual for applied research*. Sage.
13. Derek Hansen, Ben Shneiderman, and Marc A Smith. 2010. *Analyzing social media networks with NodeXL: Insights from a connected world*. Morgan Kaufmann.
14. Juan Pablo Hourcade, Sarah L Mascher, David Wu, and Luiza Pantoja. 2015. Look, My Baby Is Using an iPad! An Analysis of YouTube Videos of Infants and Toddlers Using Tablets. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 1915–1924. https://doi.org/10.1145/2702123.2702266
15. Daniel J. Hruschka, Deborah Schwartz, Daphne Cobb St.John, Erin Picone-Decaro, Richard A. Jenkins, and James W. Carey. 2004. Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research. *Field Methods* 16, 3: 307–331. https://doi.org/10.1177/1525822X04266540
16. Sun Hee Jang. 2011. YouTube as an innovative resource for social science research. In *Australian Association for Research in Education Conference (AARE 2011 Conference)*, 1–16.
17. John Lafferty and Chengxiang Zhai. 2017. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. *SIGIR Forum* 51, 2: 251–259.
18. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.
19. Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM*.
20. Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. Retrieved from https://ssrn.com/abstract=2886526
21. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, Oct: 2825–2830.
22. Rasty Turek. 2016. What YouTube Looks Like In A Day. *Medium*, 1. Retrieved January 1, 2017 from https://medium.com/@synopsi/what-youtube-looks-like-in-a-day-infographic-d23f8156e599
23. David H Wolpert. 1992. Stacked generalization. *Neural networks* 5, 2: 241–259.