Benchmark Data and Evaluation Framework for Intent Discovery
Around COVID-19 Vaccine Hesitancy

Shai Gretz1; Assaf Toledo1; Roni Friedman1
Dan Lahav1, Rose Weeks2, Naor Bar-Zeev2
João Sedoc3, Pooja Sangha2, Yoav Katz1, Noam Slonim1

1IBM Research; 2Johns Hopkins Bloomberg School of Public Health; 3New York University
{avishaig,roni.friedman-melamed,katz,noams}@il.ibm.com
{assaf.toledo,dan.lahav}@ibm.com, {rweeks,nbarzee1,psangha1}@jhu.edu
{jsedoc}@stern.nyu.edu

Abstract

The COVID-19 pandemic has made a huge global impact and cost millions of lives. As COVID-19 vaccines were rolled out, they were quickly met with widespread hesitancy. To address the concerns of hesitant people, we launched VIRA, a public dialogue system aimed at addressing questions and concerns surrounding the COVID-19 vaccines. Here, we release VIRADialogs, a dataset of over 8k dialogues conducted by actual users with VIRA, providing a unique real-world conversational dataset. In light of rapid changes in users’ intents, due to updates in guidelines or in response to new information, we highlight the important task of intent discovery in this use-case. We introduce a novel automatic evaluation framework for intent discovery, leveraging the existing intent classifier of VIRA. We use this framework to report baseline intent-discovery results over VIRADialogs, that highlight the difficulty of this task.

1 Introduction

As COVID-19 vaccines became available in late 2020, they were met with widespread vaccine hesitancy (Goldstein et al., 2015; Sallam, 2021), a phenomena recognized as a top global concern by the World Health Organization (WHO) in 2019. To address such hesitancy, one needs accurate, reliable, and up to date information, constantly available to the general public.

In recent years, task-oriented Dialogue Systems (DSs) have become an integral part of our daily lives, covering domains such as banking, tourism, and government agencies (Androutsopoulos et al., 2019).

Correspondingly, we introduced VIRA,1 the Vaccine Information Resource Assistant – an informational DS that aims to engage with COVID-19 vaccination concerns and questions. VIRA is able to respond to 181 different concerns, accumulated over the course of the pandemic. VIRA responses were written and vetted by leading medical and public health experts, relying on up-to-date facts and guidelines. An example of a dialog conducted with VIRA is presented in Table 1.

We refer to this list of concerns and questions as intents. This is a slightly modified definition of intents, which are usually general tasks or goals which the user tries to accomplish (Jurafsky and Martin, 2009). In VIRA the intents are phrased as complete sentences, e.g., “Is the vaccine safe?”, as opposed to a synthetic class name like “vac-cine_safe”, and the goal of the intent classifier is to predict the correct intent, to which VIRA responds to the user with a pre-defined message.

Advancing DSs depends on the availability of conversational datasets on which models can be trained. In domains where fact-based information is a necessity, e.g., healthcare, curating such datasets is often challenging: users in a crowdsourcing setting may not share their authentic concerns, may not represent public opinion, or may even attempt trolling. In addition, creating responses by highly-trained individuals is a very demanding process (Liu et al., 2021). Furthermore, even if one has collected data from a real-world DS, there could be limitations for making such data public.

The availability of VIRA enabled us to collect dialogs with real-world users, following word-of-mouth or social media advertising, presumably conveying genuine interest or concerns related to the vaccines. VIRA was launched in July 2021 and over the course of 10 months it accumulated over 8k conversations. We refer to this collection of conversations as VIRADialogs and release it as part of this work.2

After deploying a DS in a real-world setting, users may introduce new intents, which are not

1https://vaxchat.org
2https://vaxchat.org/research
part of the system’s predefined intents (Grudin and Jacques, 2019). VIRA’s use case represents such an extreme example where users’ intents change rapidly due to updates in guidelines and protocols, or as a response to new information (e.g., the outbreak of novel variants). Hence, we needed to frequently update and expand the set of user intents. This makes VIRADialogs a unique resource for Intent discovery methods. These methods aim to reveal such new intents from conversational logs, trying to identify the most salient new intents, which can then be reviewed and added to the DS using a human-in-the-loop process.

Existing datasets for intent classification and discovery (e.g., Larson et al. (2019)) were collected, at least in part, by showing crowd annotators queries and asking them to provide rephrases. Thus, for each query, a similar number of rephrases is collected. VIRADialogs, on the other hand, comes from a real-world use-case, and thus presumably better reflects how people communicate; the real intent distribution; and how it evolves over time – an aspect which as far as we know, is not covered by any existing data.

To directly evaluate intent discovery methods, one would need to annotate each user utterance with its gold intent, and compare this intent with the prediction of each method.

While this annotation approach is typically more precise, it is far from trivial in our real-world use-case considering the size of VIRADialogs and the high number of intents involved. Moreover, as we are dealing with rapidly changing user intents in light of new information about the virus and new guidelines, the distribution of user intents over time is not uniform, which means that manual annotation – even for a test set – would require continuous annotation over the whole time period. This makes manual annotation quite challenging.

As a practical alternative, we propose a novel retrospective evaluation paradigm which leverages the existing intent classifier of VIRA. We assume that this classifier, carefully developed over the entire relevant time period, covers most intents present in the data. Thus, we treat it as an Oracle to evaluate various intent discovery methods, independently in each month.

First, the Oracle is used to induce silver labels over the unlabeled user utterances. Next, to evaluate an intent discovery method, the same Oracle is used to classify intents predicted by this method to silver labels, enabling a fully automatic quantitative evaluation. We use this approach to evaluate various intent discovery methods on top of VIRADialogs and further share the code base to reproduce our experiments.3

To summarize, the contribution of this paper is three fold: i) We release VIRADialogs, a unique dataset of real-world human-machine conversations, reflecting COVID-19 vaccine hesitancy; ii) We propose and implement an automatic retrospective evaluation paradigm for intent discovery, relying on the availability of a high quality intent classifier; iii) We use our evaluation approach to report baseline performance of various intent discovery methods on top of VIRADialogs.

## 2 Related Work

### Benchmark Datasets and COVID-19 DSs.

Popular benchmark datasets for intent classification are also used to benchmark the task of intent discovery and were curated (at least in part) by asking crowd annotators to phrase intents suitable to a DS setting (e.g., Liu et al. (2019a); Larson et al. (2019)). Arora et al. (2020) introduce HINT3, a challenging benchmark whose test set comes from real chats in 3 domains. However, the test set contains less than 1,000 queries for each domain collected in a 15-day period, a relatively limited scope for intent discovery.

The pandemic outbreak led to the development of a few other DSs in this domain. Welch et al.3

![Table 1: An example of a dialog in VIRADialogs.](image)

| Side          | Text                                                                 | Intent                                           |
|---------------|----------------------------------------------------------------------|--------------------------------------------------|
| System        | Hey! Ask me your vaccine questions.                                  | Is breastfeeding safe with the vaccine          |
| User          | Hi. Should I be concerned about side effects of the vaccine if I’m  |                                                  |
|               | breastfeeding?                                                       |                                                  |
| System        | Perhaps check this out: while trials did not include breastfeeding   |                                                  |
|               | moms, COVID-19 vaccines are “non-replicating” vaccines. This type of  |                                                  |
|               | vaccine poses no risk for breastfed infants, so COVID-19 vaccines    |                                                  |
|               | are also safe when you’re breastfeeding!                            |                                                  |
(2020) introduce expressive interviewing – an interview style aiming to encourage users to express their thoughts and feelings by asking them questions about how COVID-19 has impacted their lives. Chalaguine and Hunter (2021) built and studied a DS specifically addressing COVID-19 vaccine hesitancy and showed that 20% of study participants changed their stance in favor of the vaccine after conversing with the system. While their motivation is similar to ours, the analyzed data is smaller and is coming from crowd annotators.

**Intent Discovery Methods.** Recent work by Rabinovich et al. (2022) introduced a fully unsupervised pipeline for detecting intents in unhandled DS logs. Utterances are encoded into vector representations, and a Radius-based Clustering (RBC) algorithm assigns each to an existing cluster, in case it surpasses a predefined similarity threshold; or use it to initiate a new cluster. The algorithm automatically selects the number of clusters, and does not enforce full partitioning of the underlying data, but rather enables outliers — instances that lay in isolation of discovered clusters. The paper suggests a method for selecting cluster representatives aimed at maintaining centrality and diversity.

Key Point Analysis (KPA) (Bar-Haim et al., 2020a,b, 2021a) proposes a framework that provides both textual and quantitative summary of the main points in a given data. KPA extracts the main points discussed in a collection of texts, and matches the input sentences to these key points. It has been shown to perform well on argumentative data, as well as in online surveys and on user reviews. To our knowledge, our work is the first to utilize KPA in the context of DSs.

### 3 The VIRA System

Users communicate with VIRA using either a web-based User Interface (UI) or a WhatsApp application. The general flow is that users enter free text expressing their questions and concerns about the vaccine, VIRA detects the intent within a pre-defined intent list, and in turn provides a suitable response, reviewed by medical experts. VIRA supports conversations in English. Below we describe VIRA’s main components.

**Profanity Classifier.** We use a dictionary to identify utterances with suspected offensive language, to which VIRA presents a generic response.

**Dialog-Act Classifier.** We classify each user input to one of the supported dialog acts. For certain dialog acts, e.g., ‘Hi’, VIRA presents a generic response. Full details can be found in Appendix A.

**Intent Classifier.** Intents representing distinct vaccine concerns were carefully curated through various means: using a Twitter analysis, reviewing audience questions in Zoom-based public forums hosted by authors' affiliated academic centers, and synthesizing web pages with frequently asked questions. The intents were defined also by taking into consideration the scientific knowledge towards the vaccine at that point. Over time, new concerns were identified by monitoring incoming queries to VIRA and eventually the list comprised of 181 intents, presented in Appendix G.

The requirement from VIRA was to provide specific answers to specific concerns, and general answers to general concerns — hence, “I am afraid the vaccine will change my DNA” and “I distrust this vaccine” required different answers, and thus were represented as separate intents, although the latter can be entailed from the former.

The intent classifier was trained on data collected from crowd annotators using the Appen platform. Annotators were presented with an intent and asked to express it in three different ways, as if conversing with a knowledgeable friend (see Section 6.1 for more details). The classifier’s top-ranked intent is selected for providing a response from the Response Database. If no intent passed a pre-defined threshold, a corresponding response is given.

**Response Database.** This database contains VIRA’s responses to intents. Each entry specifies multiple responses to a specific intent, to increase output diversity. The responses contain varying information and tone from which VIRA selects one randomly. The database was created and is maintained by experts in the field based on up-to-date facts and guidelines. All responses sought to minimize technical language and maintain brevity through a 280-character limit.

**Feedback Mechanism.** VIRA incorporates a feedback mechanism that enables users to correct the course of conversation. This feedback allows VIRA’s personnel to improve the system over time.
Table 2: Stats of VIRADialogs. Row 2 includes turns that are both free text and a feedback selection (see Appendix B), whereas row 4 indicates free text turns only.

(see more details in Appendix B).

All VIRA’s chats, including feedback selections and classifiers output, are recorded for off-line analysis, without storing identifiable information.

### 4 The VIRADialogs Dataset

VIRADialogs contains the interactions conducted with VIRA by actual users from July 2021 to May 2022. The full dialogues, as well as user feedback, predicted intents, dialog acts, and offensive language predictions are released to the research community. The data has been anonymized by masking locations, names, e-mails, phone numbers, and birth-dates, along with suspected offensive terms, using a range of regular expressions, the Profanity Classifier, and the spaCy Named Entity recognizer. In addition, we have excluded dialogues between 29-30, July 2021, in which VIRA was confronted with multiple chats containing offensive language, presumably from individuals who attempted to break the system. Stats of VIRADialogs are presented in Table 2.

### 5 Retrospective Intent Discovery Evaluation

An important contribution of this work is to show how to leverage an existing DS intent classifier – like the one described in Section 3, referred to as an Oracle – to automatically evaluate intent discovery methods over a collection of dialogs. An overview of the proposed approach is depicted in Figure 1. The underlying components are described below, using the following terminology:

**Oracle INTENTS**: The intents supported by the Oracle. **Silver LABELS**: Subset of Oracle INTENTS, induced over a given data. **Predicted INTENTS**: Intents predicted and phrased by an intent discovery method. **Predicted Oracle INTENTS**: Subset of Predicted INTENTS mapped by the Oracle to Oracle INTENTS.

#### 5.1 Inducing Silver LABELS

Given a set of unlabeled user utterances from conversational logs we randomly split it to train and test sets. The train set is used to induce Silver LABELS, while the test set is used for evaluation. The motivation of the train-test split is three-fold: (i) enabling to evaluate how consistent is the Oracle itself to ensure the emerging Silver LABELS are representative of the entire data; (ii) preserving an option to evaluate supervised intent discovery methods in future work; (iii) using the Oracle test set results to estimate upper bound test performance.

We apply the Oracle to predict (at most) one intent for each utterance in the train set. Utterances on which the Oracle confidence was below a pre-specified threshold are placed in a none cluster. Since each utterance is mapped to one intent, we obtain clusters of utterances around Oracle INTENTS. Next, we sort all clusters by their size, and define the top K ranked ones and their intent representatives as the Silver LABELS, where ranking criteria can vary (see Section 6.2 for a concrete example).

#### 5.2 Evaluation Method

##### 5.2.1 Matching Predicted INTENTS to Silver LABELS

Predicted INTENTS often cannot be matched directly to Silver LABELS. E.g., an intent discovery method might output “I don’t want to get a booster shot”, whereas the corresponding intent in the Silver LABELS would be “Will I need a booster shot?”. Assuming manual mapping is not feasible, we use the Oracle to map each of the Predicted INTENTS to – at most – one of the Oracle INTENTS, resulting in a set of Predicted Oracle INTENTS. Utterances of Predicted INTENTS which are not mapped due to low confidence of the Oracle are placed in a none cluster. Note, that in principle this set may contain Oracle INTENTS that were not selected as Silver LABELS.

##### 5.2.2 Evaluation Measures

We consider two types of measures to evaluate intent discovery methods: (a) the similarity of Predicted INTENTS to Silver LABELS; and (b) the similarity of cluster partitions generated on the test data by the Oracle and the evaluated method.

**Intent Discovery Measures**

We estimate the quality of Predicted INTENTS (PIs) using the Predicted Oracle INTENTS
6 Experimenental Setup

In this section we present a concrete implementation of the framework described in Section 5 using VIRA and VIRADialogs to automatically evaluate various unsupervised intent discovery methods.

6.1 The Oracle

For the Oracle we use VIRA’s intent classifier (Section 3), described below.

Data

For each intent amongst the final 181 intents covered by VIRA, we asked 18 Appen crowd annotators to contribute three different intent expressions, i.e., different phrasings of questions or comments by which they could have expressed the intent while chatting with a knowledgeable friend. Qualified annotators were paid on average 7.5-8$ an hour. After manual cleaning we ended up with 7,990 expressions, between 20-100 for each intent. We release this dataset as part of this work, contributing to the task of single-domain intent classification.

Model and Training

We split the intent expressions associated with each intent to train (65%), dev (8%), and test (27%) sets, with 5,169, 664 and 2,139 examples, respectively, over which we fine-tuned RoBERTa-large (Liu et al., 2019b). Full model implementation de-

Note that we collected data from crows annotators solely for training the intent model. VIRADialogs itself contains real interactions and is not crowd-sourced.

For each annotator, we calculate the BLEU score of its expressions w.r.t the intent. Annotators with score < 0.07 are determined as qualified, aiming at promoting diversity.

The data also contains a small set of 324 intent expressions, extracted manually from VIRADialogs.

https://research.ibm.com/haifa/dept/vst/debating_data.shtml
Table 3: # utterances in VIRADialogs splits for intent discovery evaluation.

| Fold | Train size | Test size | # SILVER LABELS |
|------|------------|-----------|-----------------|
| Jul-21 | 3,011 | 3,294 | 45 |
| Aug-21 | 1,169 | 1,285 | 43 |
| Sep-21 | 868 | 911 | 37 |
| Oct-21 | 718 | 747 | 34 |
| Nov-21 | 506 | 521 | 30 |
| Dec-21 | 799 | 805 | 40 |
| Jan-22 | 239 | 250 | 23 |
| Feb-22 | 212 | 220 | 18 |
| Mar-22 | 192 | 206 | 20 |

We apply to VIRADialogs filters to reduce noise and irrelevant input. We split the remaining utterances into monthly intervals, resulting in 10 data folds, and subsequently evenly split the utterances in each fold to train and test (indifferent to which dialogue utterances came from).

To reduce noise in generating SILVER LABELS, we additionally filter from the train set utterances classified with a dialog act (e.g., ‘greeting’) or as offensive, as the ratio of intents related to COVID-19 vaccines in these utterances is much smaller. We then apply the Oracle on each utterance in the train set, resulting in ORACLE INTENTS and corresponding clusters. We sort them based on their prevalence and define the top K as SILVER LABELS. In practice, we do this by accumulating the clusters until we reach a coverage of 80% (out of all texts on which the Oracle had a confident prediction) or that the number of utterances mapped to an intent is below 3 (removing a long tail of small clusters). The number of utterances and SILVER LABELS for each fold are reported in Table 3.

6.3 Intent Discovery Methods

6.3.1 Clustering Algorithms

We evaluate two clustering algorithms. Since one cannot assume that the number of SILVER LABELS is known a priori, we use \( \sqrt{N} \) as a simple heuristic to determine the number of clusters, including the none cluster, where \( N \) is the number of utterances being clustered. Short utterances, containing less than 5 recognized words, were placed in advance in the none cluster. Analysis takes a few minutes on CPU.

**K-Means.** We use the K-Means algorithm from the SciKit-Learn package (Pedregosa et al., 2011) with the default settings. Each utterance was represented using its Sentence-BERT representation (Reimers and Gurevych, 2019).

**Sequential Information Bottleneck (sIB).** As a bag-of-words baseline, we use the sIB algorithm of Slonim et al. (2002). The algorithm uses as input the Term-Frequency vector representations and is executed with the default settings, after stop-word filtering and stemming.

**Intent Extraction**

We select a single user utterance per cluster to represent an intent, resulting with the list of PREDICTED INTENTS. The selection is based on a statistical analysis of n-grams in the data. For each cluster, we first find the n-grams that are significantly more common in this cluster compared to other clusters based on hyper-geometric test \( (p = 0.05) \). Then we select the user utterance in the cluster that includes the maximal number of significant n-grams found in that cluster.

6.3.2 End-to-End Methods

We evaluate two end-to-end methods with mostly default settings. These methods determine the number of clusters internally, and map utterances to a none cluster as they see fit. For comparison purposes, we take the top sqrt(\( N \)) – 1 prevalent clusters for evaluation. The rest of the clusters are added to the none cluster.

**Key Point Analysis (KPA).** We use KPA as provided by the IBM Debater Academic Early Access Program (Bar-Haim et al., 2021b). The underlying model of KPA matches utterances with key point candidates, identified automatically. Adjustments for this task can be found in Appendix D. The service took about 3.5 hours to complete the analysis.

**Radius-based Clustering (RBC).** We approached the authors of Rabinovich et al. (2022) to produce the results for this evaluation. Adjustments for this task can be found in Appendix E. RBC took a few minutes to run on CPU.
Recall | Precision | F1 | JS-distance
--- | --- | --- | ---
0.79(±0.08) | 0.8(±0.08) | 0.8(±0.08) | 0.16(±0.04)

Table 4: Evaluation of the Oracle on VIRADialogs test sets (weighted-avg over the monthly intervals.)

7 Results and Discussion

7.1 The Oracle

We first establish the quality of VIRA’s intent classifier used as the Oracle in various ways.

Inference on Intent expressions test set. We evaluate the Oracle on the test set of the collected intent expressions, using the threshold tuned on the dev set (Section 6.1). The Oracle achieves a micro-averaged precision / recall / f1 of 0.85 / 0.74 / 0.79 on dev, and 0.88 / 0.77 / 0.82 on test.

Inducing Silver labels and matching Predicted Intents. We manually evaluate the Oracle’s accuracy in (i) inducing Silver labels (Section 5.1) and (ii) matching Predicted Intents to Silver labels (Section 5.2.1).

For (i), we randomly sample 10 Silver labels from the train set of each of the 10 folds. For each silver label we sample 2 utterances mapped to it (200 < utterance, Silver labels > pairs overall). For half of the pairs, we randomly replace the silver label with one of the other Oracle intents (thus, obtaining negative pairs). We asked 3 annotators to annotate whether a given pair of texts has a similar intent or meaning, and took the majority vote as the ground-truth (see more details in Appendix F). The accuracy of the Oracle on this data is 0.85.

For (ii), we randomly select from each fold and for each evaluated method 5 pairs of < Predicted Intents, Predicted Oracle Intents > where Predicted Oracle Intents are part of the Silver labels (200 pairs overall). We use the same annotation task as in (i). The accuracy of the Oracle on this data is 0.86.

Consistency over VIRADialogs test. To recall, we evaluate methods on the test set w.r.t. Silver labels induced from the train set. Here, we would like to examine the consistency of the Oracle’s predictions between the sets which also implies the representativeness of the Silver labels for the entire data. We do that by inferring the Oracle over the test set of each monthly fold to produce clusters around Oracle intents. We then rank them by prevalence and accumulate them to define the Predicted Intents (which are also trivially Predicted Oracle Intents), as was done to induce Silver labels on the train set. The results are presented in Table 4. The Oracle achieves a weighted-f1 of 0.795, demonstrating reasonable consistency between the train and test split in each fold. This also can be considered an upper limit of success for other methods.

Overall, the above evaluation has shown that the Oracle performs well in matching utterances and Predicted Intents to intents, and that Silver labels are relatively representative.

7.2 Intent Discovery Methods

Results for the 4 methods we evaluate are presented in Table 5. RBC has the highest coverage uncovering 45% of the Silver labels, and reaching an f1 of 0.51. These results also indicate the difficulty of this task, as the majority of Silver labels remain undetected. Note that similar precision with worse recall, such as with K-Means compared to KPA, suggests more redundancy in the Predicted Intents of the former.

KPA is much better at the clustering measures, and is thus useful for finding good examples for each intent. This might be due to KPA’s matching engine, trained to match sentences with key points (similarly to intents in VIRA, key points are concise representations of main points in the data).

It should be noted that for simplicity we used “off-the-shelf” methods with minor adaptations, to resemble a real-world setting where a user would like to get a fast impression of how well such methods perform for a given use-case with minimal effort. In addition, we used a simple heuristic to determine the number of clusters. It is likely that with proper tuning of parameters, domain adaptation of underlying models, tuning of number of clusters, etc., the performance would have been higher.

7.3 Qualitative Analysis of Emerging Intents

The Silver labels and Predicted Oracle Intents cover varying issues, and so we sought to analyze some of the more high-profile ones in light of events that occurred in their context.

We selected two intents: i) How effective is the vaccine against the Omicron variant, coupled with the rise in Omicron-related cases in December
In this paper we first describe VIRA, an information DS addressing hesitancy towards COVID-19 vaccines. VIRA provides access to accurate, up-to-date information in English, written by experts. We believe that the associated VIRA Dialogs data, containing 8k dialogs of VIRA with real-world users, would be a valuable resource to the relevant research community. As an initial example of the potential of this data, we demonstrate how it can be utilized to evaluate intent discovery methods. We propose an automatic evaluation framework that relies on the availability of a corresponding intent classifier, and report the results of 4 diverse methods, concluding that this benchmark represents a significant challenge.

While automatic evaluation is clearly more practical than manual one, developing the required intent classifier involves a non–trivial effort. Still, we envision two potential outcomes of our work. First, additional intent-discovery methods can be easily evaluated over VIRA Dialogs data using our implementation, and compared to the baseline performance reported here. Second, the same framework can be implemented in other use cases as well for which a reliable intent classifier is available, opening the door for automatic evaluation of intent discovery methods over additional datasets.

Finally, VIRA is constantly maintained and updated, and is now being expanded to additional languages, along with a Whatsapp implementation, to expand its outreach. In future work we intend to

Table 5: Evaluation of intent discovery methods on VIRADialogs. The numbers are a weighted-average over the monthly intervals. Best method for each metric is highlighted in bold. Takeaway: Methods are able to uncover up to 45% of the intents, demonstrating the difficulty of this task. RBC is able to uncover more intents and at better precision. KPA is much better at uncovering correct placements of utterances within clusters.

|          | Recall | Precision | F1  | JS-distance | AMI   | AMI   | Clustering-f1 | V-measure |
|----------|--------|-----------|-----|-------------|-------|-------|---------------|-----------|
| sIB      | 0.39(±0.09) | 0.52(±0.09) | 0.44(±0.09) | 0.33(±0.03) | 0.01(±0.03) | 0.24(±0.03) | 0.37(±0.05) |
| K-Means  | 0.44(±0.09) | 0.62(±0.09) | 0.51(±0.09) | 0.32(±0.04) | 0.15(±0.04) | 0.28(±0.05) | 0.35(±0.04) |
| KPA      | 0.44(±0.08) | 0.54(±0.09) | 0.49(±0.07) | 0.32(±0.04) | 0.24(±0.04) | 0.30(±0.04) | 0.40(±0.05) |

Figure 2: Cluster ratios of How effective is the vaccine against the Omicron variant (left); Will I need a booster shot (right). Takeaway: Predictions of methods on VIRADialogs correlate well with real-world developments.

2021, and ii) Will I need a booster shot, coupled with booster recommendations in late November 2021 and March 2022. In Figure 2, we plotted the cluster ratio of each intent among all clusters in a given month, as predicted by the Oracle, KPA, and RBC on the test set. Presumably, high ratio indicates a peak of interest for this intent.

For Omicron, methods highlight emerging interest in December and January, correlated with its real-time occurrence. To the right, methods predict interest in boosters peaking in December and April. We also note that differences between systems are sometimes non-negligible (e.g., as evident by the different peaks in the right figure). Overall, this analysis demonstrates how outstanding events in the COVID-19 timeline can be captured by the evaluated intent discovery methods.

8 Conclusions

In this paper we first describe VIRA, an informational DS addressing hesitancy towards COVID-19

---

16 https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/mathematical-modeling-outbreak.html
17 https://www.cdc.gov/media/releases/2021/s1129-booster-recommendations.html
18 https://www.cdc.gov/media/releases/2022/s0328-covid-19-boosters.html
19 Cluster ratio is defined as the size of an intent cluster divided by the overall number of utterances for a given month.
report the lessons learned from developing VIRA, and the implications for developing a DS in the public health domain.

Acknowledgements

We thank Ella Rabinovich, Roy Bar-Haim, Yoav Kantor, Lilach Eden and the anonymous reviewers for their insightful comments, and Edo Cohen-Karlik, Alex Michel and Elad Venezian for their contribution to VIRA.

9 Limitations

There are a few limitations to our approach, which stem from assumptions made to establish the evaluation pipeline.

- We implement an evaluation pipeline on a single dataset, which we were part of creating, and did not test its compliance with additional datasets.

- We assume a relatively accurate intent classifier, referred to as an Oracle, is available. Thus, our evaluation is not suited for cold-start scenarios.

- We assume the intents covered by the Oracle indeed cover most intents expressed in the data. It is quite possible that as VIRA-dials is a large dataset it included additional intents, beyond the 181 covered by the Oracle, which probably impacted the accuracy of the evaluation. We note, though, that automatic evaluation, as proposed in this work, is always prone to such issues.

- We evaluated only certain unsupervised methods for intent discovery. Other systems may perform better than the reported baselines.

10 Ethics Statement

This paper describes work around VIRA, a real-world DS addressing COVID-19 vaccine hesitancy. In an attempt to alleviate concerns that users would take action based on information given to them by VIRA which might harm them, the terms of use of the DS state that “This information ... is not intended as a substitute for medical advice”. We were guided with the principle of providing accurate information, thus when building VIRA we incorporated a direct mapping between intents and responses. Future endeavours based on this dataset, e.g., for building a generative bot for addressing vaccine hesitancy, should be aware of the ramifications of showing to users such content.

In addition, the terms of use stated that queries are stored and may be used for research purposes. The chats collected might have originally contained offensive language, often as a result of the sensitivity of the domain to some users. We made a dedicated effort to flag these cases and mask problematic terms. However, we did so with automatic measures, so the dataset might still contain such language. Finally, although the data was anonymized by masking various expressions, it is still possible that some sensitive medical concerns remain.

References

Aggeliki Androutsopoulou, Nikos Karacapilidis, Euri-pidis Loukis, and Yannis Charalabidis. 2019. Transforming the communication between citizens and government through ai-guided chatbots. Government Information Quarterly, 36(2):358–367.

Gaurav Arora, Chirag Jain, Manas Chaturvedi, and Krupal Modi. 2020. HINT3: Raising the bar for intent detection in the wild. In Proceedings of the First Workshop on Insights from Negative Results in NLP, pages 100–105, Online. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020a. From arguments to key points: Towards automatic argument summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4029–4039, Online. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Yoav Kantor, Roni Friedman, and Noam Slonim. 2021a. Every bite is an experience: Key Point Analysis of business reviews. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3376–3386, Online. Association for Computational Linguistics.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020b. Quantitative argument summarization and beyond: Cross-domain key point analysis. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 39–49, Online. Association for Computational Linguistics.

Roy Bar-Haim, Yoav Kantor, Elad Venezian, Yoav Katz, and Noam Slonim. 2021b. Project Debater APIs: Decomposing the AI grand challenge. In
Lisa A. Chalaguine and Anthony Hunter. 2021. Addressing popular concerns regarding COVID-19 vaccination with natural language argumentation dialogues. In Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 16th European Conference, ECSQARU 2021, Prague, Czech Republic, September 21-24, 2021, Proceedings, volume 12897 of Lecture Notes in Computer Science, pages 59–73. Springer.

Susan Goldstein, Noni E. MacDonald, and Sherine Guirguis. 2015. Health communication and vaccine hesitancy. Vaccine, 33(34):4212–4214. WHO Recommendations Regarding Vaccine Hesitancy.

Jonathan Grudin and Richard Jacques. 2019. Chatbots, humbots, and the quest for artificial general intelligence. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI ’19, page 1–11, New York, NY, USA. Association for Computing Machinery.

D. Jurafsky and J.H. Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall series in artificial intelligence. Pearson Prentice Hall.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Siyang Liu, Chuie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019a. Benchmarking natural language understanding services for building conversational agents.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Marina Meilă. 2007. Comparing clusterings—an information based distance. Journal of Multivariate Analysis, 98(5):873–895.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Ella Rabinovich, Matan Vetzer, David Boaz, Vineet Kumar, Gaurav Pandey, and Ateret Anaby-Tavor. 2022. Gaining insights into unrecognized user utterances in task-oriented dialog systems.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. CoRR, abs/1908.10084.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

Malik Sallam. 2021. Covid-19 vaccine hesitancy worldwide: A concise systematic review of vaccine acceptance rates. Vaccines, 9(2):160.

Noam Slonim, Nir Friedman, and Naftali Tishby. 2002. Unsupervised document classification using sequential information maximization. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’02, page 129–136, New York, NY, USA. Association for Computing Machinery.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research, 11(95):2837–2854.

Charles Welch, Allison Lahnala, Veronica Perez-Rosas, Siqi Shen, Sarah Seraj, Larry An, Kenneth Resnicow, James Pennebaker, and Rada Mihalcea. 2020. Expressive interviewing: A conversational system for coping with COVID-19. In Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, Online. Association for Computational Linguistics.

A Dialog-Act Classifier

This classifier is used for categorizing the user input as one of the supported dialog acts: greeting, farewell, negative reaction, positive reaction, concern and query. The classifier was trained on utterances extracted from early chats labeled for their dialog act. VIRA responds to input texts that are classified with one of the first 4 dialog act types with corresponding generic texts. For example, a response to a greeting (e.g., ‘Hi’) is “Hello, what
are your thoughts about the COVID-19 vaccine?”. Utterances classified as either concern or query are passed to the Intent Classifier.

B Feedback Mechanism

VIRA incorporates a feedback mechanism that gives users the option to correct the course of conversation. When users give a thumbs down for a VIRA’s response, or when the intent classifier is not confident, VIRA shows to the user the top-3 predicted intents in a menu to select from with additional options for indicating that: (a) none of these intents address the concern, or (b) the input does not express a concern at all. This feedback allows VIRA’s developers and persons maintaining the Response Database to improve the system over time. For example, when (b) is selected, it indicates a false positive for the Dialog-Act Classifier.

C Intent Classification Model Details

As a base model for fine-tuning the intent classifier of VIRA, used as the Oracle, we use RoBERTa-large (354M parameters). We use AdamW optimizer with a learning rate of 5e-6 and a batch size of 16. We fine-tune the model for 15 epochs and select the best performing checkpoint on the dev set according to overall accuracy. Training took 20 minutes on 1 v100 GPU. The confidence threshold of the model was tuned by taking the minimal threshold such that the precision on the dev set > 0.85, resulting in a threshold of 0.296.

D Key Point Analysis Details

First, utterances for which no match was found above a threshold are placed in a none cluster.

Furthermore, preliminary experiments have shown KPA is producing too few intents, so as an adjustment for this task we: (i) set limit_n_cands = false to remove the limit on the number of key point candidates; (ii) set n_top_kps = 1000 to remove the limit on number of clusters in the output, which also implies no minimal cluster size. The hypothesis is that (i)+(ii) will increase the amount and diversity of resulting key points at the expense of run-time.

E Radius-based Clustering Details

As an adjustment, chit-chat utterances which are filtered at the first phase of the algorithm are placed in a none cluster. The minimal similarity threshold is set to 0.55. As with KPA we do not set a minimum size for clusters.

F Labeling User Utterances and Predicted Intents to Silver Labels

We presented annotators with pairs of texts, where one text can be either a user utterance or an intent from the Predicted Intents, and the other a silver label. We asked, “Do the above two texts convey the same meaning or intent?”. The annotators belong to a group with high success on previous tasks of our team, and the task included a few positive and negative examples to illustrate our objective. In addition, we included test questions of text pairs manually selected from the training data of the Oracle, and annotators with less than 70% accuracy on them were removed from the task.

G Intents Supported by VIRA

| Intent | Description |
|--------|-------------|
| COVID-19 is not as dangerous as they say | |
| Do I need to continue safety measures after getting the vaccine? | |
| How long until I will be protected after taking the vaccine? | |
| How many people already got the vaccine? | |
| I am afraid the vaccine will change my DNA | |
| I am concerned getting the vaccine because I have a pre-existing condition | |
| I am concerned I will be a guinea pig | |
| I’m concerned the vaccine will make me sick. | |
| I am not sure if I can trust the government | |
| I am young and healthy so I don’t think I should vaccinate | |
| I distrust this vaccine | |
| How much will I have to pay for the vaccine | |
| I don’t think the vaccine is necessary | |
| I don’t trust the companies producing the vaccines | |
| I don’t want my children to get the vaccine | |
| I think the vaccine was not tested on my community | |
| I’m not sure the vaccine is effective enough | |
| I’m waiting to see how it affects others | |
| COVID vaccines can be worse than the disease itself | |
| Long term side-effects were not researched enough | |
| Are regular safety measures enough to stay healthy? | |
| Should people that had COVID get the vaccine? | |
| Side effects and adverse reactions worry me | |
| The COVID vaccine is not safe | |
| The vaccine should not be mandatory | |
| Do vaccines work against the mutated strains of COVID-19? | |
| They will put a chip/microchip to manipulate me | |
| What can this chatbot do? | |
| What is in the vaccine? | |
| Intent | Question |
|--------|----------|
| Are women more likely to get worse side effects than men? | |
| How do I convince my family and friends to get the COVID-19 vaccine? | |
| Why are COVID-19 vaccination rates slowing in the U.S.? | |
| I'm going to get vaccinated | |
| Is getting vaccinated painful? | |
| What do I do if I lose my COVID-19 vaccination card? | |
| Can I get swollen lymph nodes from the vaccine? | |
| Can my newborn become immune to COVID-19 if I'm vaccinated? | |
| "COVID-19 is over, why should I get the vaccine?" | |
| Did one woman die after getting the J&J vaccine? | |
| Do people become magnetic after getting vaccinated? | |
| Does the vaccine contain eggs? | |
| How is the COVID-19 vaccine different than others? | |
| How soon after I've had COVID-19 can I get the vaccination? | |
| Is it safe for my teen to get the vaccine? | |
| Is this Pfizer vaccine equally effective in kids as it is in adults? | |
| Were the COVID-19 vaccines tested on animals? | |
| What are the side effects of the vaccine in children? | |
| What is the delta variant? | |
| What is the J&J vaccine? | |
| What is the Moderna vaccine? | |
| What is the Pfizer vaccine? | |
| Where are we required to wear masks now? | |
| Who can get the Pfizer vaccine? | |
| Who can I talk to about COVID-19 in person? | |
| Why should I trust you? | |
| Will my child need my permission to get vaccinated? | |
| Will the US reach herd immunity? | |
| Will my child miss school when they get vaccinated? | |
| Is the vaccine FDA approved? | |
| Why do vaccinated people need to wear a mask indoors? | |
| Do vaccinated people need to quarantine if exposed to COVID-19? | |
| What is Ivermectin? | |
| Does the Johnson and Johnson vaccine cause Rare Nerve Syndrome? | |
| What is the difference between quarantine and isolation? | |
| Does the COVID-19 vaccine cause autism? | |
| Does the vaccine cause impotence? | |
| Who is required to get vaccinated under the federal vaccine mandate? | |
| Is the Delta variant more dangerous for kids? | |
| Will there be a booster shot for J&J and Moderna? | |
| Is the booster shot dangerous? | |
| Can I get the vaccine if I have Multiple Sclerosis? | |
| Do children receive the same dose of Pfizer as adults? | |
| What is the Omicron variant? | |
| How effective is the vaccine against the Omicron variant? | |