Diving into Robocall Content with SnorCall

Sathvik Prasad
North Carolina State University
snprasad@ncsu.edu

Trevor Dunlap
North Carolina State University
tdunlap@ncsu.edu

Alexander Ross
North Carolina State University
ajross6@ncsu.edu

Bradley Reaves
North Carolina State University
bgreaves@ncsu.edu

Abstract

Unsolicited bulk telephone calls — termed "robocalls" — nearly outnumber legitimate calls, overwhelming telephone users. While the vast majority of these calls are illegal, they are also ephemeral. Although telephone service providers, regulators, and researchers have ready access to call metadata, they do not have tools to investigate call content at the vast scale required. This paper presents SnorCall, a framework that scalably and efficiently extracts content from robocalls. SnorCall leverages the Snorkel framework that allows a domain expert to write simple labeling functions to classify text with high accuracy. We apply SnorCall to a corpus of transcripts covering 232,723 robocalls collected over a 23-month period. Among many other findings, SnorCall enables us to obtain first estimates on how prevalent different scam and legitimate robocall topics are, determine which organizations are referenced in these calls, estimate the average amounts solicited in scam calls, identify shared infrastructure between campaigns, and monitor the rise and fall of election-related political calls. As a result, we demonstrate how regulators, carriers, anti-robocall product vendors, and researchers can use SnorCall to obtain powerful and accurate analyses of robocall content and trends that can lead to better defenses.

1 Introduction

Robocalls, more precisely called "unsolicited bulk telephone calls" or "SPAM calls," have become so common that some commentators in industry estimate that they nearly outnumber legitimate calls [1]. Such a high rate of nuisance activity serves to effectively destroy the utility and trustworthiness of the global telephone network. The consequences to individuals are severe: while average users are frustrated, vulnerable users are defrauded of significant sums of money. Society pays a greater cost when a network that can connect virtually any human to any other human in real-time with unparalleled reliability becomes useless because individuals no longer answer calls for fear that they are spam.

In the United States, this state of affairs has not escaped the notice of telephone service providers, regulators, law enforcement, and legislators. While many robocalls were already illegal, in late 2019 the US Congress passed the TRACED Act. This law further enhanced penalties for illegal calls (including rampant caller ID spoofing) and mandated all providers to implement anti-robocall measures by Summer of 2021. These measures include behavioral analysis (e.g., a single account making far too many calls), blocklisting known offenders, and deploying STIR/SHAKEN, a new mechanism to authenticate the source of VoIP (Voice over Internet Protocol) calls. These mechanisms all have fundamental limitations and have not significantly reduced robocall volumes [2]. Calls that originate outside of a country cannot be completely authenticated by STIR/SHAKEN, nor can any call that ever transits a legacy phone circuit, and behavioral detection can be easily evaded by savvy operators. Finally, regulators, operators, and law enforcement are stymied both by the sheer volume of illegal calls and difficulties in collecting evidence of illegal activity. In particular, what differentiates illegal calls from legal ones is often a matter of the content, which to-date has been impossible to analyze at scale.

This paper changes the equation by introducing SnorCall, a framework for analyzing robocall content. SnorCall addresses the key challenges of deploying call content analysis at scale: limited training data, limited analyst time, unreliable content identification and classification, and incomplete understanding of the space of robocalls. It does so by leveraging recent advances in semi-supervised learning [3], word embedding and topic analysis 4.2, and natural language parsing and task identification [4, 5]. A key element of this is the Snorkel framework for semi-supervised learning [3]; Snorkel allows for rapid labeling of unlabeled data through a combination of simple but imprecise user-defined functions and a generative model that trains on those inputs. While we are not the first to study robocall content [6–10], we do claim to be the first to do so with highly accurate and repeatable methods on the largest corpus of real robocall audio to-date.

This paper makes the following contributions:

- *Design SnorCall:* We present the design of SnorCall, a framework that enables rapid development of high-accuracy models to automatically label and analyze call content. We establish a systematic codebook of robocall labels, implement five Snorkel labelers for call content, and design analyses to extract named entities and calls-to-action. We evaluate these techniques using a corpus of manually labeled data, finding labeler accuracy ranging from 90–100%.
- *Large-scale Robocall Content Analysis:* We apply SnorCall to 26,791 campaigns representing 232,723 calls over a 23 month period — the longest longitudinal study of robocalls to-date. We also are the first work to quantify the relative prevalence of Social Security Fraud, Tech Support Scams, and election-related robocalls.
- *Fine-grained Robocall Content Analysis:* We show how SnorCall analysis can reveal subtle trends in robocalling. These include an analysis of how different scam operations were affected by the COVID-19 pandemic, identifying novel twists on existing scams, and determining the median payment amount requested and the most common brands impersonated in tech support scams.
- *Infrastructure Analysis:* While caller ID can be spoofed, robocallers must own and operate the telephone numbers they instruct their targets to call. We demonstrate high-accuracy extraction of callback requests, finding that roughly half of all robocalls use them to some extent, about one in six use it as the only method of interaction. They are also shared across otherwise seemingly-unrelated campaigns. While these numbers are only used for a median of eight days, their presence in recent or historical data would still allow law enforcement to obtain the true identity of the owners — providing strong evidence to prosecute illegal calls.

2 Background and Related Work

The vast majority of robocalls feature pre-recorded audio intended to inform or persuade the listener to take some action. Some robocalls are desirable; examples include notices of school closures or public safety notifications. However, the vast majority of such calls are not desired by their recipient. In some cases, the calls are unwanted but otherwise benign and legal; examples include political messaging[1], non-profit fundraising, or telephone surveys. However, some of the most active robocalling operations commit outright impersonation or fraud [7, 11]. Other calls are sales pitches for products or services that may or may not actually exist. Operations responsible for these frauds have estimated revenues in the range of millions of dollars [12]. Robocalls remain a major problem despite extensive technical [13] and legal measures [14–16] designed to stop them. The failure of these mechanisms has a number of root causes.

Spoofing caller ID is trivial, and illegal robocallers also regularly establish new accounts with providers to continue operating. Prior research has proposed adding authentication to legacy signalling protocols [17], in- or out-of-band end-to-end authentication [18, 19], fingerprinting devices [20] or call channels [21, 22], and work on human factors [23–25]. The latest approach is a protocol called STIR/SHAKEN [26]. This protocol appends a signature from the originating provider to VoIP signalling, authenticating the call origin. While large carriers have implemented this protocol, most small carriers have not. An additional problem is that calls originating overseas cannot be reliably authenticated, and STIR/SHAKEN is incompatible with the substantial amount of legacy telephone network infrastructure.

Determining the true source of a call requires a time-consuming, manual process called *traceback*. Traceback requires the provider of the called party to identify the telephone carrier that delivered the call to the end provider. A traceback request is then sent by the called party provider to that carrier which sent the call to them. Since this intermediate carrier is usually not the originating provider, the intermediate carrier must take the same action for the carrier that sent the call to them. This process repeats recursively until the originating provider is determined. Each request must be initiated manually by a fraud detection engineer, and traceback at each hop can take a business day or longer to complete. Ultimately, this process cannot scale to the millions of robocalls placed each day. STIR/SHAKEN, if deployed widely enough, is hoped to simplify this process by allowing the terminating provider to jump straight to the originator of the call. In the meantime, providers and individuals rely on commercial products that are imperfect but effective in some cases. These products use proprietary methods likely to be similar to behavioral methods studied in the literature [10, 27–29].

Even if scalable call provenance is someday available, there are still a number of barriers to ending illegal robocalls. While with traceback it may be possible to identify the provider account of the robocaller, many commercial VoIP providers do not maintain reliable records of the true identity of the account holder. This means that an account can be closed but the culprits are free to move on to other providers for service. The FCC and many industry insiders believe most illegal robocalls originate outside their destination country, which drastically complicates enforcement of criminal or civil penalties even if a robocall operation is identified. Moreover, regulators and providers do not have staff or resources to take action – especially legal action – on every robocall, even if all of the other issues were resolved.

Most prior work on understanding robocall abuse has been limited by a paucity of data, with most work (including this paper) relying on data collected by honeypots [7, 30–32], shared by a provider [28], or captured through external reports of

[1] Political robocalls happen to be explicitly exempt from do-not-call regulations in the US.

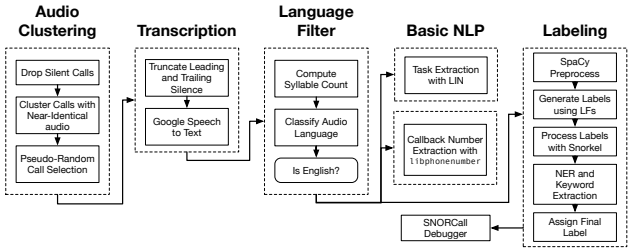

Figure 1: SnorCall comprises a five-stage pipeline of audio and transcript processing.

abuse [33–35]. Much of the prior work focuses primarily or exclusively on metadata like call records, either because call audio was unavailable or too costly to work with at scale. Prior work analyzing call content focused on a sample selection of calls [6–8] or on auxiliary data about the calls, like complaints or news reports [9, 36]. To date, two papers have used transcripts from honeypots to cluster unsolicited calls on LSI topic modeling of the transcripts [8, 10]. While these projects shared example topics and anecdotal impressions of the transcripts, the focus of the work was on estimating blocklist effectiveness, not on analyzing or characterizing call content. By contrast, because the distinction between a "good" robocall and a "bad" or illegal robocall is *semantic*, our work is focused on providing regulators, carriers, and researchers with detailed automated content analyses.

3 Audio and Transcription Processing

The SnorCall framework comprises a five-stage pipeline as shown in Figure 1. We describe the initial audio processing, transcription, and language detection stages in this section. In the next section, we describe the final two stages. The SnorCall Debugger module is described in Section 4.3.

3.1 Data Collection

We obtained call audio and call metadata (including Call Data Records, or CDRs) by operating a honeypot as described by Prasad et al [7]. Telephone honeypots consist of a set of telephone numbers that receive calls along with the infrastructure for automatically answering phone calls and storing call audio and metadata. In this paper, we study the calls placed to 5,949 telephone numbers over a 23 month period from Jan 1st 2020 to Nov 30th 2021. The honeypot answers each inbound call and plays a standard "Hello" greeting built-in to the Asterisk enterprise VoIP system. After answering and playing the greeting, the honeypot separately records inbound and outbound audio for 60 seconds, before ending the call.

The honeypot telephone numbers were donated for research by a major VoIP provider. They were donated gradually, from February 2019 through July 2019, and contain a mixture of never-assigned numbers, previously-assigned numbers ready for reassignment, and numbers taken out of use by the general public because they were either used to conduct abuse or were frequent targets of abuse (e.g., excessive robocall volume). After July 2019, all numbers used in the study were owned and operated by the honeypot. Additionally, once numbers were added to the honeypot, they were never used to place outbound calls or provided to third parties for any purpose. As a result, all calls to these numbers are by construction unsolicited.

Legal and Ethical Considerations: Both legal counsel from our university and our IRB were thoroughly briefed on the data collection procedures and subsequent analyses, and both offices approved the data collection and analysis. Our IRB determined that our research protocol as submitted was exempt from human subjects research review. Legally, the primary concern is if consent from both call parties is required for lawful recording. The honeypot is operated in the United States, and both US federal law and the law of the state where the honeypot is operated allow either party to the call to record the call without confirming consent.

Ethically, the main concerns involve issues of respect for persons because it is possible that a live human calls the honeypot. This concern motivated numerous design choices in honeypot operation. The honeypot is operated such that any caller faces negligible risks compared to their normal activities. Calls are limited in duration to prevent any consequences from keeping a line open too long (e.g., preventing other calls to the caller). The audio clustering step (described in Section 3.2) ensures that we only study calls with multiple instances of repeated audio, meaning that an occasional accidental call will be ignored in our analysis. We do not take steps to traceback or positively identify the caller through any means, protecting their identity. Further, we are under a nondisclosure agreement to not release raw data or other content that could potentially identify a caller (e.g., their phone number). When using a third party service for transcription, we paid a premium to ensure that our audio would not be shared or used for training later editions of the speech-to-text model. Finally, no actions were taken by the honeypot operators to "seed" or otherwise encourage calls, meaning that all calls arrived voluntarily on the part of the caller.

3.2 Audio Processing

Audio Clustering: Robocallers by definition reuse audio recordings in their calls, so we use an audio clustering pipeline to group robocall audio recordings based on very high audio similarity. The pipeline processes recordings on a monthly basis to study the evolution of campaigns over our study period. The resulting clusters correspond to robocall campaigns with near-identical call scripts. The audio clustering is similar to the approach described by Prasad et al. [7]. The high accuracy of this pipeline assures that calls within a cluster are virtually

identical. However, some robocalls dynamically insert the target's name within the call audio. In such cases, the pipeline may split a campaign into multiple campaigns based on the extent of variation within the audio. To conserve transcription resources, we process one randomly-selected call from each audio cluster using the transcription pipeline.

Transcription: The transcription pipeline converts a representative robocall audio recording from each campaign into text using Google's Speech-to-Text (STT) online transcription service specifically designed for phone call audio. We truncate leading and trailing silences in recordings for a substantial cost reduction without loss of information.

Before using Google's STT service, we tested the quality of transcripts generated by multiple online and offline transcription frameworks. We recorded and transcribed audio files from speakers with multiple accents for these tests. We also used public robocall recordings in English and non-English languages from FCC's public awareness webpages. Google's STT service generated superior quality transcripts with minimal development effort. Google's phone call transcription service also ranks high on transcription benchmarks [37]. Despite our efforts to identify the best transcription framework, and significant advancement in Deep Learning based Speech-to-Text frameworks [37], there are occasional errors in phone call transcripts. These errors are expected due to the inherent lossy characteristics of audio traversal through the phone network. Our results and claims are cognizant of such transcription errors throughout the paper.

3.3 Language Detection

Most robocalls within the North American Phone Network use English to communicate with the call recipient. However, a small yet significant fraction of robocalls use Spanish, Mandarin and other non-English languages [7]. Because well-known NLP techniques and libraries are language specific, we create a language detection pipeline to filter out non-English robocalls (including Fax and Modem tones).

The key insight behind the language detection pipeline is that the transcript length is proportional to the amount of *English* content within a call when using an English language transcription service. A long call in English with significant amount of audio and minimal silence will generate a long transcript, while a long call with significant amount of audio in a different language will generate a short, incoherent English transcript. We verified this behavior of the transcription service by transcribing 10 robocalls in Spanish and Mandarin.

Using this insight, we build a classifier to identify English calls by using non-silent audio duration and transcript syllable counts as inputs. We created a dataset of randomly selected audio samples across three languages: English, Mandarin and Spanish [38, 39]. Our choice of languages were based on evidence from prior research [7], which reported that Mandarin and Spanish robocalls are prevalent in North America. For

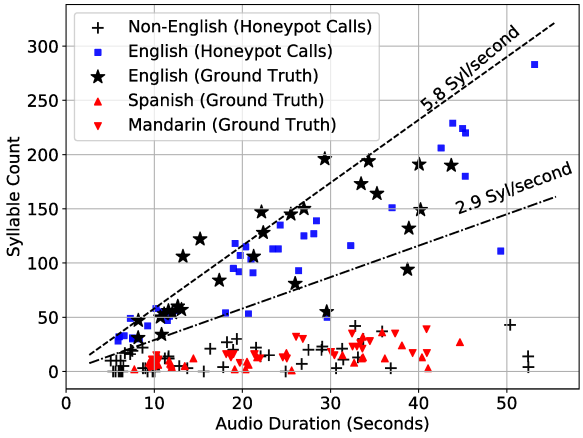

Figure 2: Classifying English and Non-English calls

English samples, half of the samples were from male speakers and the other half were from female speakers. The dataset already contained information about the speaker's gender. For non-English samples, the dataset did not contain information about the speaker's gender. These samples included background noise. We made sure that the final samples had a variety of audio durations (0-15, 15-30 and 30-45 seconds). We repeated the same steps for Spanish and Mandarin audio samples. We re-sampled the labeled dataset audio recordings to 8kHz to ensure that they matched the characteristics of the audio collected in the honeypot. Using WebRTC VAD, we computed the amount of audio in each sample (Feature 2). To test the model, we selected 10 English and 10 non-English calls from the honeypot data by manually listening to the audio. We transcribed each audio sample using Google's Speech-to-Text phone call transcription model (US English, 8kHz sample rate, with automatic punctuation). We computed the number of syllables from the audio transcripts [2].

We trained a ridge classifier from scikit-learn — a linear classifier — with 3-fold cross validation and achieved a mean accuracy of 0.97. This indicates that our approach of using audio duration and syllable count to classify English and non-English audio recordings is effective. Figure 2 shows how English and non-English calls group together when considering syllable count and audio length.

4 Semantic Analysis

The previous section explained how we obtained reliable English-language transcripts from robocalls. In this section, we discuss how we apply NLP techniques to these transcripts to extract semantic information about these calls. We begin by describing the process by which we developed a comprehensive codebook of topics for robocalls. We then describe how we automatically assign labels from this codebook to

[2] *Syllables*: https://pypi.org/project/syllables/

transcripts, and how we evaluate our analysis. The section concludes with a discussion of how we extract "calls-to-action" from transcripts.

4.1 Robocall Codebook

Before developing techniques to label robocalls, we undertook an effort to develop a codebook of potential robocall topics as a guide for what labels would be useful to implement in SnorCall. A robocall labeling codebook was essential for two reasons: (i) we needed a systematic approach to categorize the transcripts by understanding the semantics of the call content and (ii) a labeling codebook enabled us to manually label data, which was used to evaluate our models.

The four authors collaboratively developed the codebook using deductive and inductive approaches. The authors reviewed existing research on robocalls [7, 13, 35, 40], public reports from regulatory agencies, and derived insights from professional experience with working on robocall detection and mitigation. We found that the robocall categories described in prior work were far from comprehensive given the vast variety of robocalls observed by victims and phone users. Further, we found that none of the prior work considers useful robocalls like district-wide school notifications and public safety robocalls. For completeness and to simplify organization, we followed the structure of the latest available North American Industry Classification System (NAICS) [41]. The NAICS is the standard used by the US Economic Census to hierarchically classify both public and private sectors of the US economy, so we can be assured that our codebook covers virtually any economic activity. We refined the codebook iteratively by labeling multiple sample datasets of robocall transcripts and by meeting regularly to discuss improvements to the codebook. The final codebook consists of 7 top-level categories and 33 sub-categories, where some sub-categories are further split into smaller categories. The codebook can be found in Appendix B.

4.2 Labeling Robocalls with Snorkel

The core of SnorCall is an automated robocall labeling pipeline using Snorkel [3], a semi-supervised data labeling framework. With most predictive modeling, a large amount of manually labelled data is required to achieve reasonable success. This data often comes at great labor expense. By contrast, Snorkel relies on humans to define a number of simple, lightweight *labeling functions* (LFs) to assign preliminary labels to otherwise unlabeled data. The Snorkel framework is built on the assumption that the outputs from individual labeling functions are noisy, imprecise, and possibly correlated. For certain inputs, labeling functions can even return a label that is contrary to the label a human expert would assign to the input robocall transcript. The Snorkel framework expects that the individual accuracies of labeling functions are unknown and accounts for correlated labeling functions. The decision to assign a specific label to the input robocall transcript is performed by a label aggregation [42, 43] stage using a generative model within the Snorkel framework. The implementation of this generative model, called `LabelModel`, is based on Ratner et al.'s work [42] on aggregating the output from multiple weak supervision sources (labeling functions). The `LabelModel` learns the accuracy of individual Labeling Functions, computes weights for individual labeling functions based on training data, and assigns the final label for individual data points in the prediction phase. We describe additional details about LF development process in Section A.1 and discuss the keyword extraction process in Section A.2.

Training the `LabelModel`: A small number of robocall transcripts annotated by human experts were used to train the `LabelModel`. Before we developed a new Snorkel for a particular label, a robocall expert identified a set of keywords and phrases that were indicative of a robocall belonging to that label/category. For example: for a Snorkel for POLITICAL label, the keywords used were `vote, election, campaign, trump, biden`. Similarly, for Social Security Snorkel, the keyword was `social security`. Using the keywords defined for each Snorkel as a search phrase, we randomly sampled two equal sized sets of 300 robocall transcripts — (i) a positive set which contains the keywords and (ii) a negative set which does not contain any of the keywords.

Our choice of training dataset size was based on the experiments and guidelines of the authors of `LabelModel` [42] and Snorkel [3]. In the foundational paper for `LabelModel` [42], Ratner et al. showed significant performance on three different classification problems with training datasets with only 200–350 samples.

A human expert manually reviewed each transcript to make sure that they belonged to the appropriate set — positive or negative. If there was a conflict, the transcript was discarded and the process was repeated to generate two equal sized positive and negative sets. This process was repeated to develop a test dataset of at least 10% the size of the training set for each Snorkel. This test dataset was used to iteratively improve the performance of SnorCall during the training process, and not for Snorkel performance evaluation described in Section 4.3.

Iterative Snorkel Development: We followed the guidelines set by the authors of the Snorkel framework. We evaluated the performance of our labeling functions, iteratively fine tuned them to maximize the performance of each Snorkel pipeline [44]. As part of this iterative process, we studied three performance metrics for each Labeling Function: Coverage, Overlaps and Conflicts. *Coverage* is the fraction of training data points on which at least one LF has returned a NON-ABSTAIN label. *Overlap* is the fraction of training data points with more than one NON-ABSTAIN label. Among the overlapping data points, *Conflict* is the fraction of training data points where the Labeling Functions disagree on. We followed Snorkel development best practices to improve the

coverage, and minimize conflicts and overlap [44].

Labeling Robocall Transcripts: Robocall transcripts are labelled with one or more labels using a combination of labeling functions and the respective generative model (`LabelModel`). Each robocall transcript passes through the set of Labeling Functions to compute the input matrix. This input matrix is fed into the trained generative model. The `LabelModel` can perform one of the following two options: *(i) Assign a label:* If the transcript is being processed by a Snorkel pipeline that labels POLITICAL calls, and if the `LabelModel` predicts that the input transcript is a political robocall, a POLITICAL label is assigned to it. *(ii)* ABSTAIN *or assign a negative label:* If the political Snorkel pipeline cannot establish that the input is a political robocall, it may chose to ABSTAIN from assigning a label or may assign a negative label that indicates that the input is not POLITICAL. At the output phase, our framework treats ABSTAIN output as a negative label.

Snorkel Implemented: For time and space reasons, we developed 5 Snorkels for this paper using the techniques listed above. We chose a selection of topics to explore abuse topics from prior work (Social Security and Tech Support Scam calls), explore the effects of election robocalling (Political calls), and topics that may or may not indicate fraud or abuse (Financial and Business Listing calls).

4.3 Ground Truth and SnorCall Evaluation

We ensured that the ground truth dataset used to evaluate SnorCall contains robocall samples from a wide range of robocall categories. We sampled 300 transcripts after grouping similar transcripts together based on word similarity. Even though the ground truth dataset consists of 300 transcripts, they collectively represent 2,490 individual robocalls, since each transcript represents a robocalling campaign uncovered using the Audio Clustering pipeline from Section 3.2. This approach captured a broad spectrum of robocalls from different categories in the ground truth data, while being conscious about the time-consuming task of manually labeling each robocall. Three authors independently labeled these 300 transcripts and resolved conflicts by discussing with each other.

To manage and label transcripts, we used NVivo, an application designed to help researchers label data. We labeled multiple rounds of sample datasets to become accustomed to the labeling environment, the codebook, and the overall labeling process. Each author used the codebook discussed in Section 4.1 and assigned one or more appropriate leaf nodes from the codebook. Depending on the SnorCall type, a parent level node was used to aggregate all the transcripts present under that parent node. Using this process, a ground truth data set for Social Security, Political, Financial, Tech Support and Business Listing was developed. We process the 300 transcripts using all five SnorCall models and provide the evaluation results in Table 1. The Social Security SnorCall had a 100% accuracy, since it was able to correctly identify the 7 Social Security robocalls in the ground truth data while correctly recognizing the remaining 293 as non-Social Security robocalls. In Section 6, we discuss how SnorCall's system performance impacts our results.

Table 1: Snorkel Evaluation Results on Ground Truth Data

Snorkel	Precision	Recall	Accuracy
Social Security	1.00	1.00	1.00
Tech Support	0.72	0.87	0.98
Financial	0.73	0.71	0.90
Political	0.70	0.78	0.95
Business Listing	0.64	1.00	0.98

SnorCall Debugger: In a landscape where robocallers adopt new strategies based on societal events, SnorCall models will need to be re-trained to account for concept drifts [45]. Like most classifiers, SnorCall models have a tradeoff between precision and recall. As shown in Figure 1, SnorCall includes a debugger module that enables domain-experts to address these concerns. It highlights the named entities recognized in a transcript, extracts the internal training weights and votes (positive, negative, or abstain) of individual LFs during inference, and provides the probability value for the final label assignment. During the iterative training and prediction phase, this information is used to inspect False Positives and False Negatives, explain the final label assignments in terms of individual LFs, update LFs, or retrain SnorCall models when appropriate. A sample output from the debugger module is shown in Figure 11.

Challenges of working with robocall data: Unlike processing coherent text documents using NLP techniques, analyzing raw audio collected over the phone network poses a distinct set of challenges. Real-world audio recordings are prone to noise and data loss (packet loss, jitter, and delay). We methodically developed multiple pre-processing stages in our pipeline (silence detection, audio de-duplication, and language detection) to improve our data quality for downstream analysis. Despite these efforts, some transcripts contain mismatched capitalization, incoherent punctuation, grammatical and other minor errors. We observed sub-par performance for specific NLP tasks (e.g. NER) while processing noisy transcripts. This was not surprising because NLP models are often pre-trained on coherent text, and robocall transcripts are inherently noisy due to the lossy nature of the phone network.

4.4 Call-to-Action in Robocalls

Robocall originators intend to invoke one or more actions from call recipients [35, 40]. Some example actions include "...pressing one to talk to an agent", "...call us at ...", "...donate now to our campaign" etc. In this paper, we refer to such action verbs as *call-to-action* verbs. The benefit from studying the call-to-action within a robocall is that we

can provide insights on how robocallers intend to interact with the call recipients.

We extract call-to-action verbs from robocall transcripts by applying state-of-the-art task extraction techniques from the NLP literature. We process transcripts using *Lin* [5] a domain-independent task extraction framework. Lin extracts tasks by identifying the main verb in a sentence using syntactic (dependency parsing and linguistic rules) and semantic features of the sentence. We used Lin because it performs well on unknown domains [5], outperforms baseline techniques on task extraction and could be tailored [3] to our use-case.

We found that 81.52% (20,549) of all campaigns had at least one task. Each task is represented as a tuple of verb phrase and an object, eg. "(press, one), (visit, us)". Lin successfully extracted 3,231 such (verb, object) tuples with 669 unique verbs. However, some of these tuples identified by Lin are not necessarily a valid call-to-action or instruction to the call recipient in the context of a robocall, eg. ('forget', 'everything'). We group tasks based on the verb and manually review the tasks. During manual analysis, we identify the (verb, object) tuples that can be used during a conversation between a caller and the call recipient to indicate a call-to-action. By following this manual process of validating the (verb, object) usage, we identified 131 unique verbs that indicate a call-to-action within a robocall. 72.79% (18,348) campaigns used one of the 131 verbs as a call-to-action.

5 Results

5.1 Data Characterization

Call Volume and Dates: Our 23 month study started on Jan 1st 2020 and ended on Nov 30th 2021. Over this duration, the honeypot answered 1,355,672 calls across 5,949 phone numbers. Throughout the data collection process, the honeypot had an approximate downtime of less than 10 days for infrastructure maintenance and unexpected power outages.

Audio Clustering: After performing silence detection preprocessing, 371,045 (27.37%) calls contained sufficient audio information for further analysis. The Audio Clustering stage described in Section 3.2 uncovered 26,791 monthly campaigns consisting of 232,723 (17.17%) calls in total, with an average of about 1,165 campaigns each month. A monthly campaign seen in the honeypot contained 8.69 calls on average. During each month, the honeypot observed campaigns with hundreds of calls with some campaigns containing thousands of calls. These outlier campaigns reinforce that robocalling operations reuse audio recordings or use largely similar audio recordings to generate bulk calls over long periods of operation. Further, these outlier campaigns demonstrate the effectiveness of the Audio Clustering stage in Figure 1 to reliably uncover large groups of similar audio recordings consistently.

[3]Lin: Task extraction https://github.com/Parth27/Lin

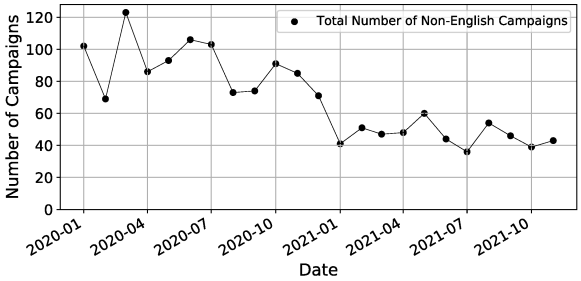

Figure 3: Hundreds of Non-English campaigns were active throughout the study period

Audio and Silence Duration within Robocalls: The preprocessing and audio clustering stages collectively uncovered robocalling campaigns with dense audio. Each English campaign consisted of calls that were on average 32.19 seconds long. About 68.72% of the call contained audio, with the rest being silence. The significant percentage of silence within robocalls can be attributed to the call originator waiting for the user to perform an action or respond to the audio prompt. This finding demonstrates the importance of trimming silence from the leading and the trailing ends of a robocall recording collected in the honeypot. Trimming silence substantially reduced the transcription cost.

Transcription Statistics: We transcribed 26,791 representative calls into text — each call representing a campaign. The average length of a transcript was 64.41 words. An average call duration of 32.19 seconds indicates that the caller generally engages in a conversational rate of speaking at about 2 English words per second [46].

Language Detection: The Language Detection stage described in Section 3.3 uncovered 25,206 (94.08%) English campaigns and 1,585 (5.92%) Non-English campaigns over our study period. The English campaigns consisted of 226,488 calls (97.32%) and the Non-English campaigns consisted of 6,235 calls (2.68%). Even though the number of Non-English campaigns are multiple orders of magnitude lower than English campaigns, the perennial characteristics of Non-English campaigns observed in Figure 3 are noteworthy. Spanish and Mandarin robocalls constantly attempt to target the non-English speaking populations [7] in North America.

5.2 Individual Label Analysis

This section describes our findings from each robocall category analyzed by SnorCall.

5.2.1 Social Security Robocalls

Finding 1: *SnorCall uncovered 1,304 (5.17%) Social Security campaigns consisting of 8,292 (3.66%) calls in total*

during the study period. These scams impersonate the Social Security Administration, a federal government agency, and threaten call recipients with dire consequences in an attempt to steal sensitive information like their Social Security Number. The following sample transcript from one of the social security campaign demonstrates persuasion tactics of fabricated threats and a false sense of urgency, which are frequently reported by victims [47].

> This call is regarding to your social security number. We found some fraudulent activities under your name and arrest warrant has been issued and your Social Security would be suspended soon. Please press one to talk with officer right away. I repeat, please press one to talk with officer right away. Thank you.

In each of the 1,304 campaigns, the caller impersonated the Social Security Administration using at least one of the numerous variations of the term "Social Security Department". Some example variations were: "The Department of Social Security", "Social Security Administration", "SSA Department", etc. Through SnorCall's ability to identify named entities, we automatically identified multiple variations, and correctly labeled these references as organizational names.

Finding 2: *Social Security scams have moved to targeting Social Security Disability*. The robustness of SnorCall helped us uncover a lesser known type of Social Security scam targeting the vulnerable population that seeks Social Security benefits. We observed two broad variants of Social Security scam calls operating over multiple months. The first type was the well-known scam where the impersonator threatens the target and persuades them to reveal their SSN using a false sense of authority and scarcity [47]. However, SnorCall uncovered a new variant of Social Security scam calls where the callers were impersonating Social Security disability advisors. These calls seem well-intended and non-intimidating. The caller establishes a sense of prior commitment to persuade the target to respond using a false sense of authority [47]. Consider the following transcript:

> Hello, my name is Audrey and I'm a social security disability advisor with national disability on a recorded line. And my call back number is XXX-XXX-XXXX. Now I show here that you recently inquired about your eligibility for Social Security disability benefits. Can you hear me? Okay?

This variant of the Social Security scam call is not well-known [47], with limited online public awareness information [48]. This lesser known variant comprised of 515 campaigns with a total of 3,498 calls. The more popular type of scam calls impersonating the Social Security Administration consisted of 789 campaigns with a total of 4,794 calls, as shown in Figure 4.

Finding 3: *The operational characteristics of Social Security scam calls changed substantially during the onset of the COVID-19 pandemic*. As many countries started imposing local restrictions and lockdowns due to the COVID-19 pandemic, there was a substantial decline in the number of Social Security scam calls in the honeypot, as seen in Figure 4. Even

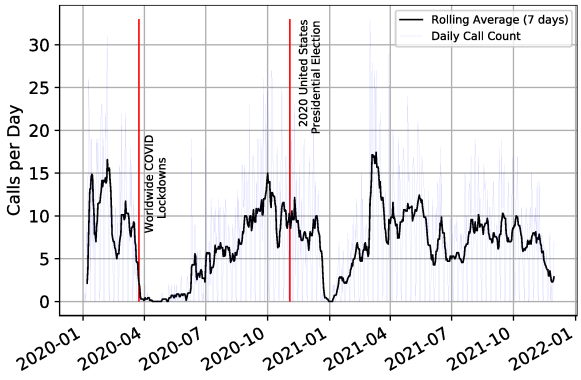

Figure 4: Calls impersonating the Social Security Administration were impacted by worldwide COVID-19 lockdowns and reduced substantially during the Christmas and New Year's break in 2021

though we do not have evidence to explain why we observed a substantial decline in Social Security scam calls during the COVID-19 lockdowns, anecdotal evidence from the Scam baiting community [49] suggests that some of the Social Security scammers were operating from dedicated work locations. Since most COVID-19 related restrictions prohibited people from commuting to work and to other places outside their homes, the lockdowns directly affected business operations that required people to work from offices and other designated locations. We speculate that this drastic decline in the call volume indicates that Social Security scam operations were disrupted by lockdowns, and were operating from dedicated work locations with office-like infrastructure.

Finding 4: *Social Security scams mention other government entities while impersonating the Social Security Administartion in an attempt to increase the credibility of the caller*. We found references to the US Treasury (20 campaigns, 73 calls), the Federal Bureau of Investigation (FBI) (11 campaigns, 57 calls), the Drug Enforcement Agency (DEA) (3 campaigns, 6 calls), and the Federal Trade Commission (FTC) (1 campaign, 5 calls). Many of these campaigns were operating over multiple months. This finding supports prior research [47] in which victims of Social Security scam reported that scammers bolstered their credibility by falsely claiming the involvement of additional federal agencies like the FBI, DEA, DoJ (Department of Justice), etc. When victims questioned the legitimacy of the call from Social Security, scammers often initiated an additional call or a three-way conference call, falsely demonstrating the involvement of these additional federal agencies [47]. Overall, SnorCall's capability to identify numerous Named Entities helped us glean deeper insights into the operational strategies of Social Security scam calls and other calls.

> From us Drug Enforcement Administration the reason of this call to suspend your social insurance number on an immediate basis as we have received suspicious Trails of information with your social security number the moment

> you receive this message. You need to get back to us to avoid legal consequences on *XXX-XXX-XXXX*. I repeat *XXX-XXX-XXXX*. Thank you.
>
> This call is from legal Department of Social Security Administration. The purpose of this call is regarding an enforcement action, which has been executed by the US Treasury against your social security number ignoring. This would be an intentional attempt to avoid initial appearance before the Magistrate Judge or exempt jury for federal criminal offense. If you wish to speak to social security officer now, please press one. Thank you.
>
> Is to inform you that Social Security Administration is suspending your social security number because of the criminal activities and money laundering frauds done by you the investigation team of our department and the FBI are investigating more about this case File. We are trying to reach you for the last time to verify about such activities just in case we will not hear back from Thursday will be considered as an intentional fraud and the lawsuit will be file under your name in order to get more information and talk to the social security officer. Kindly, press one. I repeat, press one to connect your call to social security officer. Thank you.

Finding 5: *Lack of direct references to money or a dollar amount in Social Security scam calls indicate that the initial robocall is the beginning of an elaborate sequence of events to engage the target*. Among 1,304 campaigns, only 7 campaigns had a direct reference to money through a monetary value. Absence of direct mention of a dollar amount indicates that Social Security scams attempt to scare the victims by threatening them with arrests or other dire consequences [47]. As reported by victims of Social Security scams [47], once the target engages with the caller, the perpetrators employ social engineering tactics and elaborate deception techniques to deceive their target and cause them financial harm.

5.2.2 Tech Support Robocalls

Technical Support scams or Tech Support scams are a class of fraudulent operations in which a malicious entity impersonates a technical expert from a well-known technology company and defrauds the victim with the intent of causing financial harm. Such scams are prevalent across platforms (social media platforms like Twitter, YouTube, Snapchat, etc., email and the telephone network). Social media platforms with dedicated content moderation teams regularly monitor such content and remove them. However, telecom companies find it challenging to monitor and take timely action to block such operations when they operate over the phone network.

Finding 6: *SnorCall uncovered 2,696 (10.70%) tech support campaigns consisting of 8,402 (3.71%) calls*. While much of the consumer awareness and popular media on robocalls has focused on auto warranty, Social Security, and IRS scams, tech support scams remain a less-known but still formidable threat to citizens. SnorCall isolated such robocalls and enabled us to study a wide variety of fraudulent behavior employed by tech support scammers.

Prior work [12] suggests that a common strategy among tech support scams is to impersonate well-known consumer technology companies. Such scams also reference products (Figure 5) or services (Figure 6) that are associated with that consumer technology company to capture their target's attention. Using SnorCall, we were able to extract the Named Entities of products and organizations, study the behavior of tech support robocalls and analyze how such calls impersonate technology companies and reference popular products.

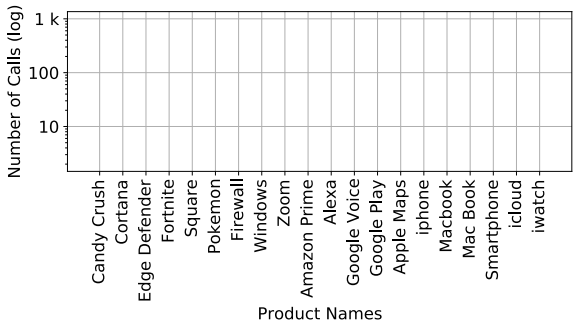

Figure 5: Apple products were frequently referenced in tech support robocalls, along with other online services

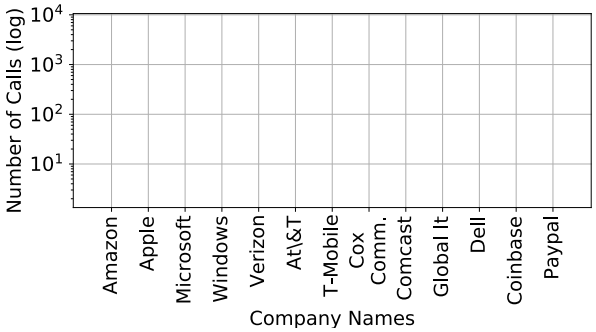

Figure 6: Tech Support robocalls reference well-known consumer-facing tech companies and less common companies like phone carriers

Finding 7: *We found that tech support scammers impersonate a wide range of consumer technology and services companies*. Prior work by Miramirkhani et al. [12] extensively analyzed tech support scams focussing on Microsoft Windows computers and other Windows utilities. We find that the volume of Amazon tech support scam calls are multiple orders of magnitude greater than well-known Windows tech support scams. Such changes in impersonation strategies is a crucial indicator on how tech support scammers adapt over time. Figure 6 shows the distribution of top 16 companies referenced by tech support scam calls with the largest being Amazon (1,477 campaigns, 7,046 calls). Consider these example transcripts:

> John from Amazon customer service. My employer ID is AMC, 2516, and I'm calling you about your Amazon Prime account. I wanted to inform you that your Amazon Prime account is being compromised, as long as an order for an iPhone 10 worth $549. For which the card attached to your Amazon account is being charged. We have placed, hold on it. As that order seems to be

> fraudulent. So please call us on this toll free number XXX-XXXX. I repeat one *XXX-XXX-XXXX* to talk to an Amazon fraud department executive. Thank you.

> Purchase from Amazon shopping. This call is to inform you that your purchase for Apple MacBook Pro and apple airpods will be delivered by tomorrow evening and $1,537.35 Home in debited from your account for this purchase. If you authorize this charge then no action required, and if you did not authorize this charge, then press one to connect to Amazon customer representative for cancellation charge.

> Dear Verizon Wireless Customer your account has been suspended for verification to reactivate your account. Please press one to speak with a customer service representative, please press two.

Three variations of the Amazon tech support calls stood out to us — (i) call originators claiming to represent Amazon's fraud department citing a discrepancy on the victim's Amazon account, (ii) call originators warning the victim about automated renewal of their Amazon Prime membership and offered assistance to remediate the charges, and (iii) call originators claiming to be an Amazon associate warning the user about a product order (Apple iPhone, MacBook, etc). Interestingly, there were also numerous calls impersonating wireless cellular carriers — AT&T (38 campaigns, 301 calls), T-Mobile (2 campaigns, 6 calls), and Verizon (6 campaigns, 14 calls).

> Dear coinbase. Customer your coinbase account, temporarily disabled indicates that your account currently has a restriction, potentially related to a security concern. This restriction requires a coinbase Security review to be removed. This restriction, may be applied for several reasons. Our security team suspected that your account was being targeted by a malicious user. Please. Press one to contact customer support for recover, your Bitcoin, please press one for recover your Bitcoin.

Finding 8: *We uncovered 3 campaigns consisting of a total of 6 calls where the caller claimed to represent Coinbase customer support*. These calls impersonated Coinbase support agents claiming that the call recipient's account was locked. As seen in the sample transcript above, the caller urged the call recipient to press a number on their keypad to receive assistance in recovering the locked Bitcoin. With the growing popularity and adoption of mobile cryptocurrency wallets and exchange platforms, impersonating cryptocurrency platforms is a lucrative strategy for tech support scammers. To achieve high success-rate for cryptocurrency scams, robocallers would have to specifically target the phone numbers of cryptocurrency wallet users. However, by designing scams based on popular services like Amazon or well-known products like Windows or MacBook computers, tech support scammers can target a much larger population [50] by indiscriminately calling random phone numbers.

Among tech support scam campaigns, we encountered certain campaigns offering Search Engine Optimization (SEO) services for Google search results [51]. Even though such SEO calls impersonate Google and offer technical assistance for website hosting, these calls specifically target small businesses and not individuals. Therefore, we exclude SEO related calls while analyzing impersonation of consumer technology companies.

As seen in examples of tech support transcripts in Section 5.2.2, the impersonators mention the dollar value of products or services being offered as part of the scam. For example, the requested value may refer to the approximate cost of phones, laptops, anti-virus subscription services, gift cards, etc. In some cases, the scam is intentionally vague. They describe a flagged transaction and an associated dollar value without referencing a specific product or a service.

Finding 9: *The median amount requested in a tech support call is about $400*. Using SnorCall, we extracted all references to Money as a named entity. We discarded non-numeric reference to a dollar value, for example "a couple of hundred dollars", "at least a few hundreds dollars", etc. On studying the distribution of the filtered dollar values, we encountered outliers which seemed unreasonable. We manually listened to the audio recordings of such robocalls and found that some tech support campaigns were using poor quality Text-to-Speech systems to state the dollar value. The poor quality pronunciation resulted in transcription errors, and in-turn skewed the distribution of dollar values mentioned in the call. We identified $1,539 as the maximum threshold value by manually listening to calls in the descending order of the dollar amount and discard all values beyond $1,539. After preprocessing and discarding outliers, we found that the median return on conversion value for tech support scam calls is $399.99 and the mean is $513.18.

Finding 10: *Online consumer-facing services are common victims of tech support scams*. Traditional computer hardware, browsers and operating systems have been a common choice for tech support scams. However, as shown in Figure 5, we also observed names of online services, smartphones, smart devices, smart watches, and gaming platforms being used to entice victims into engaging with the caller. Consumers of such products and services are susceptible to falling victim to scams and often sustain substantial financial loss. These consequences tarnish the brand value of the organization being impersonated. SnorCall enables consumer-facing services to actively monitor their brand names and warn their customers to mitigate the impact of such impersonation scams.

5.2.3 Political Robocalls

Finding 11: *We identified 1,226 (4.86%) political robocall campaigns consisting of 11,727 (5.18%) calls during our study period*. Using robocalls to communicate with the public is a common and legal strategy employed by political candidates and political organizations. The prevalence of political robocalls in the honeypot indicates that robocalls continue to be a common means of disseminating political information in the United States. A few types of political calls among the campaigns uncovered by SnorCall were those that urged voters to cast their votes to a specific candidate or a politi-

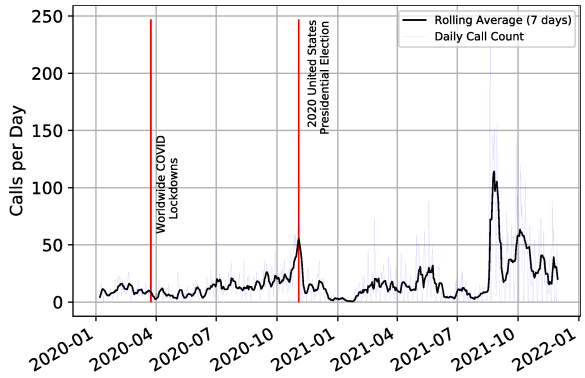

Figure 7: Large robocalling campaigns misrepresented political events from August to November 2021

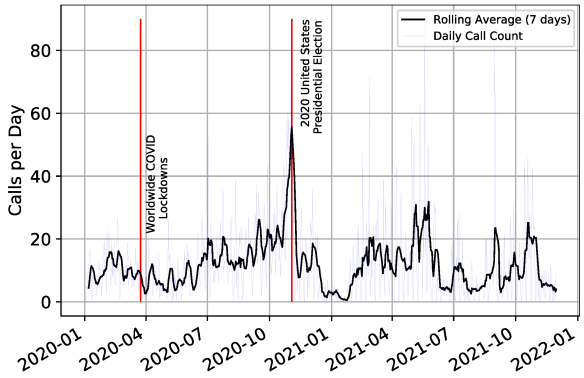

Figure 8: Political calls to the honeypot increased substantially towards the 2020 US Election Day and dropped drastically after the Election Day

cal party at the federal, state or local elections, calls seeking donations or other forms of support, and surveys calls to understand how their constituents would vote. Our data collection overlaps with the 2020 United States presidential election, which helped us glean insights into the political robocalling landscape towards the election day and beyond.

Finding 12: *Eight large campaigns misrepresented political events to drive sales or information theft*. The temporal distribution of call volume of political calls seen in the honeypot indicated a substantial increase in political calls from August to November of 2021, as seen in Figure 7. We studied the largest campaigns during these months and identified the outlier campaigns responsible for the rise in call volume between August and November of 2021. Eight outlier campaigns containing a total of 3,491 calls with 436 calls on average within each campaign were responsible for the drastic rise in political calls between August and November of 2021 in the honeypot. These campaigns were labeled as both POLITICAL and FINANCIAL, showing how SnorCall can provide multifaceted perspectives on call transcripts.

Through manual analysis of the transcripts, we discovered that the call scripts were misrepresenting political current events to encourage participation in loan or warranty schemes. Several campaigns claimed to represent a non-existent "Economic Impact Student Loan Forgiveness Program recently put into effect by the Biden Administration" and encouraged callers to enroll. While proposals for student loan forgiveness were being publicly discussed at the time, no such program was established during the campaign's activity. Given that the claimed program did not exist, it is unclear if victims will be offered loan products or if their personal information will simply be stolen. Another campaign using a similar script replaced loan forgiveness with references to "the Biden administration's infrastructure bill" that the caller fraudulently claimed provided subsidies for auto warranties that the caller was offering. Beyond demonstrating the flexibility of SnorCall, these findings also demonstrate that campaigns can adapt to current events.

Finding 13: *The volume of political calls gradually increased and reached its highest volume towards the 2020 Presidential Election day (3rd November 2020) followed by a steep decline*. We expect political parties and candidates to maximize their voter outreach efforts as they approach the election day. This includes reaching out to voters via robocalls. As highlighted in Figure 8, political campaigns identified by SnorCall captures the increase in the volume of political robocalls as we approach the election day, and the steep decline that follows. While these results may have been expected, SnorCall is able to quantify the phenomenon.

Challenges in Performing Partisan Analysis: We did not perform a partisan analysis of political calls observed in the honeypot, i.e. attempt to classify political calls into either Democratic, Republican or Independent. We note that the amount of semantic context embedded within political calls to determine the political orientation cannot be extracted with high confidence automatically, using NLP techniques. We observed that political calls regularly reference their opponents during the call, which makes it even more challenging to accurately perform a partisan analysis.

99% of all names of people referenced in political calls were the names of current or past politicians. Intuitively, a political robocall would reference people with a political background. We compared the list of unique references to a person's name in political robocalls against a publicly maintained list of political names. This regularly updated list [4] contains the names of current and past members of the U.S. Congress, congressional committees, current committee members, current and past presidents and vice-presidents of the United States using a fuzzy matching approach. With 99% of all names extracted from political robocalls matching a name in this list of political names, we have additional confidence that SnorCall was able to accurately identify political calls and extract people's names effectively.

[4] https://github.com/unitedstates/congress-legislators

5.2.4 Business Listing Robocalls

Finding 14: *We identified 1,260 (4.99%) business listing campaigns consisting of 24,316 (10.74%) calls*. Business listing calls target small and medium business owners. These calls offer online marketing or search engine optimization (SEO) on platforms like Google, Bing, Amazon Alexa network etc. They take advantage of the fact that small business owners are often not aware of online marketing and SEO techniques. Often, such claims of influencing the search results and business listing on online platforms are not legitimate [51]. Consider the following transcripts:

> Our records show that you have not updated your free Google Business listing, press one now to verify and update your Google listing, press 9 to be removed from the list again, press 1 to verify and update your Google listing.

> In Florida. I'm reaching out to you today because your company is not registered with Amazon Alexa's voice system Amazon currently has customers looking for your type of services in your area Amazon Alexa am currently in over 100 million households and is still growing, please, press one to see if your business qualifies. You can start receiving these clients immediately again, please press one to speak with our business support agents now, if you are not interested, please press two.

5.2.5 Financial Robocalls

Finding 15: *SnorCall uncovered 4,638 (18.40%) Financial robocalling campaigns consisting of 57,839 (25.54%) calls*. These campaigns were operational throughout our study period. Interestingly, we observed multiple large outlier campaigns in the months of January and February of 2020. Each campaign contained hundreds of calls offering student loan forgiveness. However, these student loan robocalls to the honeypot substantially reduced in volume over the next few months. The average call volume of financial calls remained stationary after the initial decline in the student loan forgiveness robocalls.

We observed financial robocall campaigns that advertise work/earn from home schemes, often with the promise that a large sum of money could be earned in a short amount of time. As seen in the example transcript below, the caller uses social proof and likeness as persuasion tactics to engage with the targets. Such offers were particularly enticing for people who were transitioning from an in-person job to a remote job due to the COVID-19 pandemic.

> Please stop what you're doing and listen to this short message because it could truly make a difference in your financial situation in record time or cash flow system has grown steadily for over eight years, but now due to the current restrictions. Our business is exploded and because of this unprecedented growth. We have more people now than ever bringing in $10,000 or more every 10 to 14 days using are done for you system. Listen during these trying times working from home has now become a necessity instead of just a desire. So press three months right now. If you want to find out exactly how to put $10,000 or more in your pocket every ten to fourteen days. I guarantee you've never seen anything like this up until now, so, press three right now. I get all the details or press nine and you will never hear from us again.

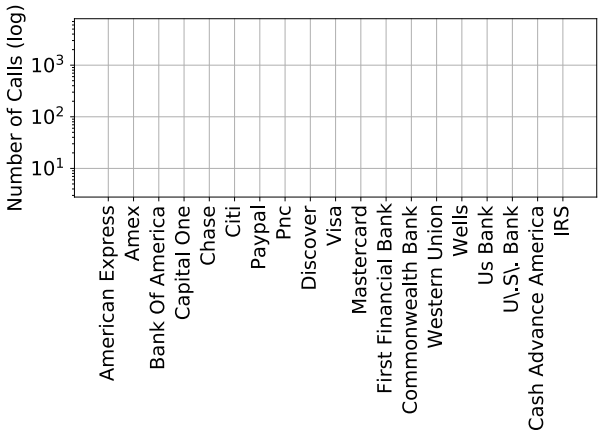

Figure 9: Financial robocalls reference well-known banks, payment platforms, credit card vendors and the IRS

Anecdotally, fraudulent financial robocalls impersonate well-known banks, credit unions and credit bureaus. Similar to tech support calls in Section 5.2.2 which impersonate technology companies, fraudulent financial calls claim to represent banks or other financial institutions to deceive people.
Finding 16: *Fraudulent financial robocalls impersonate banks, credit card vendors and federal agencies in the United States*. SnorCall identified multiple organizations referenced within financial robocalls, as shown in Figure 9. Since the category of financial calls is quite broad, it was not surprising to uncover a long list of organizations being referenced. Unsolicited calls claiming to represent well-known banks are seldom legitimate. We manually reviewed the list of organizations referenced within financial calls, and identified well-known banks and other financial organizations. We found that credit card vendors — Mastercard, Visa and Discover — were the most frequently referenced financial organizations.
Finding 17: *"Tax relief" companies use robocalls to advertise their services to taxpayers in distress*. We manually examined transcripts referencing the IRS, as shown in Figure 9. These calls advertized services which claim to eliminate tax debts, stop back-tax collection and offered solutions to reduce tax payments [52]. Historically, public radio stations and television channels have been popular mediums to advertise such offerings. SnorCall enabled us to study robocalls as a medium to advertise potentially dubious services to tax payers.

5.3 Comparing Calls-to-Action

Most robocalls contact callers with the intention of invoking an action or a response. We refer to such actions as *call-to-action* verbs and extract them using our approach described in Section 4.4. From Figure 10, we find that two common calls-to-action used by robocallers across all categories we have

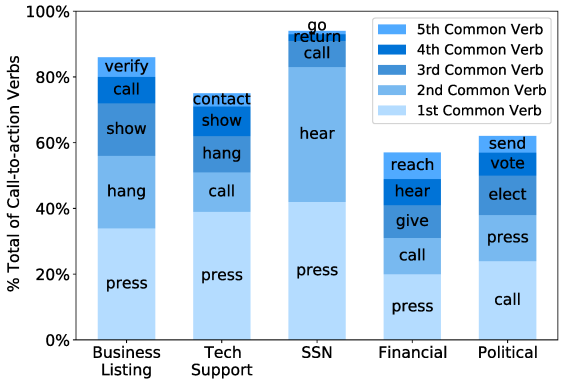

Figure 10: Instructing call recipients to press a digit is the most common call-to-action among 4 out of the 5 categories

analyzed are "call" and "press". Both actions indicate that the caller expects the target to engage with the call originator based on the information conveyed within the call. The calls-to-action in political calls — "elect" and "vote" — capture the caller's intent to engage with their constituents.

5.4 Callback Number Extraction

In this section, we describe the callback number extraction and validation process, and characterize the callback numbers embedded within the robocall audio recordings. A callback number is a phone number that a robocaller shares with the recipient after the call is answered. Callback numbers are embedded within the call audio and in-turn can be extracted from audio transcripts. Through these callback numbers, the robocaller attempts to prompt the target to initiate a phone call and potentially speak with a live representative for the robocalling campaign, as seen in example transcripts in Section 5.2.1.

Our callback number extraction process has an accuracy of 100%. Google's Speech-To-Text service recognizes spoken phone numbers and converts them to a NANP structure (i.e., NPA-NXX-XXXX) during transcription. To extract these numbers from transcripts, we use a set of regex rules to identify NANP phone number structures with the library called libphonenumber [5] to validate the number structures. We validated our callback number extraction process by randomly sampling 50 transcripts that were flagged by our extraction process to contain a callback number and 50 transcripts that did not contain a callback number. We manually reviewed the raw transcripts and compared it to the corresponding callback number extracted. We confirmed that all callback numbers were either properly extracted from the transcript when present or not extracted when not present.

Finding 18: *About 45% calls contained a callback number.* Substantial presence of callback numbers within call transcripts indicate that callback numbers are a popular engagement strategy among robocallers. Unlike asking users to press a number on their keypad after answering the call, providing a callback number is an effective engagement strategy in situations where the user does not answer the call and redirects it to voicemail. It is in the interest of the call originator to own the callback phone number and embed an active (in-service) number within the call audio so that the recipients call back and potentially speak with a live representative. Therefore, callback numbers are directly linked to the actual account holder of the robocalling campaign. About 84% of the time, callback numbers were mentioned only once during the call, and were observed in the last 10% of the message.

Finding 19: *17% calls used callback numbers as the only engagement strategy to interact with the call recipient.* These calls did not contain any other call-to-action described in Section 4.4. Among other calls with at least one call-to-action, the most common call-to-action used to engage the caller was by asking users to "press" or "dial" a digit on their phone. Engaging the call recipients through a key press or urging them to call back using the callback number embedded within the audio allows robocallers to operate independently of the asserted caller ID. If a downstream service provider starts blocking calls based on the asserted caller ID of the campaign, robocallers can change the caller ID or spoof a different caller ID. With minimal effort, they can resume their operation by reusing the same audio recording.

Finding 20: *Callback numbers rarely matched the asserted caller ID.* Among the calls with a callback number, only 4.23% of them matched the respective asserted caller ID. While legitimate reasons may exist for using different caller ID and callback numbers, it can be cause for concern. 59% of callback numbers were toll-free, allowing the recipient to call with no charge. The owners of toll-free numbers incur the cost for such calls. We speculate that legitimate campaigns are willing to take the cost burden away from the caller, while potentially malicious robocallers are willing to incur the cost of owning toll-free numbers to pose as legitimate entities.

Finding 21: *Different robocalling campaigns are sharing infrastructure.* We observed 3225 unique callback numbers from English transcripts. Compared with the number of campaigns identified, the fewer unique callback numbers indicate shared infrastructure across various campaigns. To identify shared infrastructure, we compared the unique callback numbers across campaigns. 881 callback numbers were used in more than one campaign, and 2344 callback numbers were used in only one campaign. For example, we saw one robocalling campaign selling health insurance and another contacting about a car's extended warranty; both campaigns used the same callback number. These shared callback numbers across multiple campaigns may indicate shared infrastructure being used to operate multiple robocalling campaigns.

Finding 22: *Callback numbers tend to be short-lived, with a median lifespan of 8 days.* Callback number lifespan represents the length of time the same callback number has ap-

[5] https://github.com/daviddrysdale/python-phonenumbers

peared across campaigns. The short lifespan of callback numbers implies that they have a short shelf life. Investigative leads originating from the threat intelligence gathered using callback numbers should be acted upon swiftly.

6 Discussion and Future Work

We discuss how SnorCall enables stakeholders to combat illegal robocalls, describe how SnorCall's system performance affects our results, and highlight directions for future work.

- SnorCall can automatically analyze thousands of robocalls, allowing investigators and regulators to focus on more detailed analyses. Investigators currently rely on manual analysis, where they listen to thousands of robocall recordings [53] collected via honeypots, through subpoenas or from other sources. SnorCall serves as an investigative tool to process evidence about fraudulent robocalling operations. Government agencies and consumer-facing companies that are frequent targets of impersonation scams (Findings 2, 4, 7 and 8) can monitor variants of well-known scams. Government agencies like the Social Security Administration (SSA) can use SnorCall to proactively uncover lesser-known variations of impersonation scams (like the disability scam described in Finding 2). SSA can warn Social Security disability beneficiaries about such emerging threats through consumer awareness initiatives.
- SnorCall empowers regulators and enforcement agencies to proactively uncover malicious robocalls and prioritize the takedown of egregious robocalling operations. Currently, these entities depend on reports from victims to investigate the source of illegal robocalls. The investigative task of tracing a phone call to its source is called *traceback* [54,55]. Tracebacks are time-intensive manual processes spanning days or weeks, often involving coordination between fraud detection teams across multiple carriers. Investigators can prioritize tracebacks for the more egregious scams (Social Security, IRS) over deceptive marketing calls (car warranty, business listing). Timely detection and prioritized resource allocation may minimize harm to the public.
- Carriers can use SnorCall to monitor active malicious robocalls targeting their subscribers. They can proactively engage the respective upstream carrier responsible for the malicious traffic. Carriers can also develop temporary containment strategies to block calls originating from caller-IDs that are part of the malicious campaign.
- SnorCall's callback number extraction capability, as described in Section 5.4, allows investigators to track down the entity responsible for originating illegal robocalls (Findings 18, 19, 21 and 20). Investigators can identify the organization/individual who owns the callback numbers through number ownership databases and subpoenas. They can take legal action against the call originator [53] and the carrier harboring such entities [56].

Impact of SnorCall's performance on our results: We ensure that any conclusions we draw are based only on accurate classification. While reporting our findings in Section 5.2, multiple authors manually reviewed the corresponding transcripts to ensure that the raw transcripts supported our claims. We manually verified the variants of Social Security campaigns (Findings 2, 4, 5), confirmed that tech-support calls were impersonating consumer technology companies (Findings 7, 8, 10), reviewed potentially dubious references to banks and the IRS (Findings 16, 17), and manually reviewed the campaigns misrepresenting political events (Finding 12). We also listened to the raw audio to substantiate our claim (Finding 9) when we were not satisfied with the transcripts.

Future Work: We intend to further study the use of semantic triage to assist in tracebacks. Studying SnorCall's deployment in an active investigative setting (e.g. provider's fraud team) could lead to valuable insights on the evolution of robocalling operations. Lessons learnt from such studies can help us develop heuristics to detect malicious campaigns in the early stages of its lifecycle. Understanding how non-experts use SnorCall can help us develop a user-friendly interface for SnorCall.

7 Conclusion

In this paper, we have seen how SnorCall can accurately extract semantic content from robocall audio, and how that information can inform operators, regulators, law enforcement, researchers, and the general public about robocall operations. While many of the findings were interesting in their own right, some of them — such as the call back analysis — may become essential tools in combatting illegal bulk calling.

8 Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments. We would also like to thank Bandwidth Inc. for their support and for providing VoIP service and phone numbers for the honeypot. This material is based upon work supported by the National Science Foundation under grant number CNS-1849994 and CNS-2142930. This paper was partially supported by funds from the 2020 Facebook Internet Defense Prize. This material is based upon work supported by the Google Cloud Research Credits program with the award GCP19980904. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, other funding agencies or financial supporters.

References

[1] E. Brown. Over half of the calls people receive now are spam. https://www.zdnet.com/article/over-half-of-the-calls-people-receive-now-are-spam/, April 2020.

[2] Reported fraud losses increase more than 70 percent over 2020 to more than $5.8 billion. https://www.ftc.gov/news-events/news/press-releases/2022/02/new-data-shows-ftc-received-28-million-fraud-reports-consumers-2021-0.

[3] A. Ratner, S. H. Bach, H. R. Ehrenberg, J. A. Fries, S. Wu, and C. Ré. Snorkel: rapid training data creation with weak supervision. In *International Journal on Very Large Data Bases*, 2020.

[4] spaCy. https://spacy.io.

[5] P. Diwanji, H. Guo, M. Singh, and A. Kalia. Lin: Unsupervised extraction of tasks from textual communication. In *Conference on Computational Linguistics*, 2020.

[6] M. Relieu, M. Sahin, and A. Francillon. "Doing being" an ordinary human callee. In *International Institute for Ethnomethodology and Conversation Analysis*, 2019.

[7] S. Prasad, E. Bouma-Sims, A. K. Mylappan, and B. Reaves. Who's Calling? Characterizing Robocalls through Audio and Metadata Analysis. In *USENIX Security Symposium*, 2020.

[8] A. Marzuoli, H. A. Kingravi, D. Dewey, and R. Pienta. Uncovering the Landscape of Fraud and Spam in the Telephony Channel. In *IEEE International Conference on Machine Learning and Applications*, Dec 2016.

[9] Q. Zhao, K. Chen, T. Li, Y. Yang, and X. Wang. Detecting Telecommunication Fraud by Understanding the Contents of a Call. *Cybersecurity*, 2018.

[10] S. Pandit, R. Perdisci, M. Ahamad, and P. Gupta. Towards Measuring the Effectiveness of Telephony Blacklists. In *Network and Distributed System Security*, 2018.

[11] Marriott international files lawsuit to combat fraudulent robocalls. https://news.marriott.com/news/2021/05/18/marriott-international-files-lawsuit-to-combat-fraudulent-robocalls.

[12] N. Miramirkhani, O. Starov, and N. Nikiforakis. Dial One for Scam: A Large-Scale Analysis of Technical Support Scams. In *Network and Distributed System Security*, 2017.

[13] H. Tu, A. Doupé, Z. Zhao, and G. Ahn. SoK: Everyone Hates Robocalls: A Survey of Techniques Against Telephone Spam. In *IEEE Symposium on Security and Privacy*, 2016.

[14] Telemarketing sales rule. https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/telemarketing-sales-rule.

[15] FCC. TRACED ACT or FS.151 - Pallone-Thune Telephone Robocall Abuse Criminal Enforcement and Deterrence Act. https://www.congress.gov/bill/116th-congress/senate-bill/151.

[16] M. Sahin and A. Francillon. On the Effectiveness of the National Do-Not-Call Registries. In *Workshop on Technology and Consumer Protection*, May 2018.

[17] H. Tu, A. Doupé, Z. Zhao, and G. Ahn. Toward Authenticated Caller ID Transmission: The need for a Standardized Authentication Scheme in Q.731.3 Calling Line Identification Presentation. In *ITU Kaleidoscope: ICTs for a Sustainable World*, 2016.

[18] B. Reaves, L. Blue, and P. Traynor. AuthLoop: End-to-End Cryptographic Authentication for Telephony over Voice Channels. In *USENIX Security Symposium*, 2016.

[19] B. Reaves, L. Blue, H. Abdullah, L. Vargas, P. Traynor, and T. Shrimpton. AuthentiCall: Efficient Identity and Content Authentication for Phone Calls. In *USENIX Security Symposium*, 2017.

[20] H. Sengar. VoIP Fraud: Identifying a Wolf in Sheep's Clothing. In *ACM Conference on Computer and Communications Security*, 2014.

[21] V. A. Balasubramaniyan, A. Poonawalla, M. Ahamad, M. T. Hunter, and P. Traynor. PinDr0p: Using Single-Ended Audio Features to Determine Call Provenance. In *ACM Conference on Computer and Communications Security*, 2010.

[22] S. Nagaraja and R. Shah. VoIPLoc: passive VoIP call provenance via acoustic side-channels. In *ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2021.

[23] I. Sherman, J. Bowers, K. McNamara Jr, J. Gilbert, J. Ruiz, and P. Traynor. Are You Going to Answer That? Measuring User Responses to Anti-Robocall Application Indicators. In *Network and Distributed System Security*, 2020.

[24] S. Pandit, J. Liu, R. Perdisci, and M. Ahamad. Applying deep learning to combat mass robocalls. In *IEEE Security and Privacy Workshops*, 2021.

[25] S. Pandit, K. Sarker, R. Perdisci, M. Ahamad, and D. Yang. Combating robocalls with phone virtual assistant mediated interaction. In *USENIX Security*, 2023.

[26] Secure Telephone Identity Revisited (STIR). https://datatracker.ietf.org/wg/stir/charter/.

[27] D. Ucci, R. Perdisci, J. Lee, and M. Ahamad. Towards a practical differentially private collaborative phone blacklisting system. In *Annual Computer Security Applications Conference*, 2020.

[28] J. Liu, B. Rahbarinia, R. Perdisci, H. Du, and L. Su. Augmenting telephone spam blacklists by mining large cdr datasets. In *ACM ASIA Conference on Computer and Communications Security*, 2018.

[29] H. Li, X. Xu, C. Liu, T. Ren, K. Wu, X. Cao, W. Zhang, Y. Yu, and D. Song. A Machine Learning Approach to Prevent Malicious Calls over Telephony Networks. In *IEEE Symposium on Security and Privacy*, May 2018.

[30] P. Gupta, M. Ahamad, J. Curtis, V. Balasubramaniyan, and A. Bobotek. M3AAWG Telephony Honeypots: Benefits and Deployment Options. Technical report, 2014.

[31] P. Gupta, B. Srinivasan, V. Balasubramaniyan, and M. Ahamad. Phoneypot: Data-driven Understanding of Telephony Threats. In *Network and Distributed System Security*, 2015.

[32] M. Balduzzi, P. Gupta, L. Gu, D. Gao, and M. Ahamad. MobiPot: Understanding Mobile Telephony Threats with Honeycards. In *ACM ASIA Conference on Computer and Communications Security*, 2016.

[33] S. Gupta, D. Kuchhal, P. Gupta, M. Ahamad, M. Gupta, and P. Kumaraguru. Under the Shadow of Sunshine: Characterizing Spam Campaigns Abusing Phone Numbers Across Online Social Networks. In *ACM Conference on Web Science*, 2018.

[34] P. Gupta, R. Perdisci, and M. Ahamad. Towards Measuring the Role of Phone Numbers in Twitter-Advertised Spam. In *ACM ASIA Conference on Computer and Communications Security*, 2018.

[35] H. Tu, A. Doupé, A. Zhao, and G. Ahn. Users Really Do Answer Telephone Scams. In *USENIX Security Symposium*, 2019.

[36] H. E. Bordjiba, E. B. Karbab, and M. Debbabi. Data-driven approach for automatic telephony threat analysis and campaign detection. In *Digital Investigation*, 2018.

[37] F. Dernoncourt, T. Bui, and W. Chang. A framework for speech recognition benchmarking. 2018.

[38] Kaggle. Speech samples of English, German and Spanish languages. https://www.kaggle.com/toponowicz/spoken-language-identification.

[39] openslr: Mandarin data, provided by Beijing Shell Shell Technology. http://www.openslr.org/33/.

[40] M. Sahin, M. Relieu, and A Francillon. Using Chatbots Against Voice Spam: Analyzing Lenny's Effectiveness. In *USENIX Usable Privacy and Security*, SOUPS 2017.

[41] NAICS. Naics codebooko webpage. https://www.census.gov/naics/?58967?yearbck=2017.

[42] A. Ratner, B. Hancock, J. Dunnmon, F. Sala, S. Pandey, and C. Ré. Training complex models with multi-task weak supervision. In *AAAI Conference on AI*, 2019.

[43] A. J. Ratner, C. De Sa, S. Wu, D. Selsam, and C. Ré. Data programming: Creating large training sets, quickly. In *Neural Information Processing Systems*, 2016.

[44] snorkel. June 2019 workshop. https://www.snorkel.org/blog/june-workshop.

[45] G. I Webb, R. Hyde, H. Cao, H. L. Nguyen, and F. Petitjean. Characterizing Concept Drift. *Data Mining and Knowledge Discovery*.

[46] ncvs. ncvs. https://ncvs.org/research/.

[47] Marguerite DeLiema and Paul Witt. Mixed Methods Analysis of Consumer Fraud Reports of the Social Security Administration Impostor Scam. Oct 2021.

[48] Social Security. https://www.ssa.gov/scsm/.

[49] Jim Browning. The Inside Man: The SSA Scam. https://www.youtube.com/watch?v=b9n2TX-86Vk.

[50] Cormac Herley. Why do Nigerian Scammers Say They are from Nigeria? In *WEIS*. Berlin, 2012.

[51] Google Isn't Calling You. https://www.consumer.ftc.gov/blog/2018/05/google-not-calling-you.

[52] FTC. Tax Relief Companies. https://consumer.ftc.gov/articles/tax-relief-companies.

[53] FCC Fines Telemarketer $225 Million for Spoofed Robocalls . https://www.fcc.gov/document/fcc-fines-telemarketer-225-million-spoofed-robocalls.

[54] D. Frankel. Senate Hearing on Combating Robocall Fraud. https://www.aging.senate.gov/imo/media/doc/SCA_Frankel_7_17_19.pdf, 2019.

[55] US Telecom ITG Report 2019. https://www.ustelecom.org/wp-content/uploads/2020/01/USTelecom_ITG_2019_Progress_Report.pdf.

[56] The Department of Justice Files Actions to Stop Telecom Carriers Who Facilitated Hundreds of Millions of Fraudulent Robocalls to American Consumers. https://www.justice.gov/opa/pr/department-justice-files-actions-stop-telecom-carriers-who-facilitated-hundreds-millions.

A Appendix

A.1 Developing Custom Labeling Functions

In manual labeling, a human expert uses their domain knowledge and intuition to categorize a robocall into a particular category. For example, a robocall impersonating the Social Security Administration or referencing Social Security Numbers would be labeled as Social Security scam robocall by a human labeler because of the presence of the keyword `Social Security Administration` or `Social Security Number`. Similarly, a robocall discussing achievements of a political candidate, mentioning the names of a political candidate and urging the call recipient to vote for the candidate would be labeled as a political robocall by human experts. Labeling functions are designed to capture these sorts of rough (but often effective) heuristics programmatically.

In the case of Snorkel, LFs are implemented using simple Python functions. To define an LF, the developer must first provide preprocessed data. In our case, we use SpaCy, the leading Python library for NLP. In our implementation, SpaCy's NLP pipeline uses `en-core-web-trf-3.1.0` model and consists of 6 stages: `transformer, tagger, parser, attribute_ruler, lemmatizer, ner`. Next, the developer specifies an *operation* such as determining presence or absence of a keyword, presence or absence of a Named Entity, or measuring the sentiment of the transcript. With this in place, the developer specifies a label to be returned on success; if the LF fails, then the LF returns a sentinel value ABSTAIN to indicate no information from the LF.

Snorkel supports *bipolar* LFs, which can assign a label for presence and a label for absence of a linguistic feature, as well as *unipolar* LFs that return only a single label or ABSTAIN. SnorCall decouples the bipolar labeling functions into unipolar labeling functions because of characteristics specific to the problem domain of labeling robocalls. In many situations, a negation of the *operation* does not necessarily indicate that the input should be assigned the opposite label by the labeling function. For example, the presence of an ORG tag (indicating the mention of a company name) may be an indication of a Tech Support robocall. However, an absence of ORG tag does not necessarily indicate that the call is not a Tech Support call. By decoupling bipolar labeling functions into unipolar labeling functions, we allow fine grained control to effectively translate the human expert's domain knowledge into labeling functions within the robocall labeling framework.

SnorCall offers a range of custom LFs that empower domain experts to capture and translate their insights into linguistic LFs which can label robocalls. Each robocall labeling framework's LF is built using elements of Snorkel labeling functions. We describe each LF in the Robocall Labeling Framework below:

- `NERTagPresence`: Returns the user specified `Return Label` if a particular SpaCy NER tag (eg: ORG, MONEY) is present in the input. Otherwise, returns ABSTAIN
- `NERTagAbsence`: Returns the user specified `Return Label` if a particular SpaCy NER tag (eg: ORG, MONEY) is absent in the input. Otherwise, returns ABSTAIN
- `SentimentChecker`: Returns the user specified `Return Label` if the input has a sentiment polarity specified between a lower and an upper sentiment polarity threshold.
- `KeywordPresenceExactMatch`: Returns the user specified `Return Label` if at least one of the keyword from a list of keywords are found. Otherwise, returns ABSTAIN.
- `KeywordAbsenceExactMatch`: Returns the user specified `Return Label` if all of the keywords from a list of one or more keywords are absent. Otherwise, returns ABSTAIN.
- `NamedEntityFuzzyMatch`: Returns the user specified `Return Label` if a particular SpaCy NER tag is present, and the value of this Named Entity is a close match to at least one of the values from a list of keywords/names. Otherwise, returns ABSTAIN. For example, this type of LF can be used to check for the presence of PERSON SpaCy NER tag and compare the value with a list of popular names (names of politicians). Another example could be when we would like to check for ORG SpaCy NER tag and check if the ORG is present in a subset of financial institutions (banks, credit card vendors or credit bureaus)

All the Snorkel models described in this paper are built using the labeling function types listed above. In the Social Security example stated earlier, a simple `KeywordPresenceExactMatch` LF that searches for the keyword `social security` is an example of translating human insight into a Python Labeling Function that returns a `Social_Security` label on matching the keyword. By developing a set of Labeling Function types specially designed for robocalls over the existing labeling function framework of Snorkel, we drastically reduced the time required to develop Snorkels for SnorCall.

A.2 Automatically Extracting Keywords

We further reduced development time through supervised automated keyword selection through topic modeling with BERTopic [6]. BERTopic uses a series of sBERT based sentence embeddings, performs data normalization and aggregates similar transcripts together using HDBSCAN clustering algorithm. By specifying input parameters to BERTopic, we extract coherent topics that are unigrams (single word) and bigrams (two words). We can then pass these topics directly to our keyword LFs to specify their labeling criteria. Before using the topic keywords extracted by BERTopic to generate Labeling Functions, we review them manually to ensure that they are coherent. We prune vague and generic keywords, if any. This process drastically simplifies the effort for a human to specify labelling functions.

[6] https://github.com/MaartenGr/BERTopic

A.3 Additional Figures

Figure 11: Example output of the SnorCall Debugger Module for a correctly classified political survey robocall

B CodeBook

1. **Information (51-NAICS)**
   1. Tech support calls (scams)
      1. Amazon tech support scam
      2. Google tech support scam
      3. Apple and iCloud tech support scam
      4. Microsoft tech support scam
      5. Generic PC support scam
      6. Malware or Antivirus removal scam
      7. Other tech support scam
   2. Business listing/search engine optimization (SEO)
      1. Google business listing
      2. Amazon Alexa listing
      3. Yellow pages listing
      4. Other business listing calls
2. **Finance and Insurance (52-NAICS)**
   1. Financial calls
      1. References to well-known banks (Bank of America, Chase, PNC, Morgan Stanley, etc.)
      2. References to payment apps and payment ecosystems (Venmo, Zelle, PayPal, etc.)
      3. Student loan
      4. Gift cards
      5. Credit cards (Visa, Master Card, Discover etc)
      6. Debt collectors
      7. Investment or get rich quick schemes
      8. Work/earn from home schemes
      9. Credit score improvement/credit bureau/modify credit history
      10. Lottery, prizes and sweepstakes
   2. Other financial robocalls
      1. Insurance, warranty and protection plans
      2. Health insurance
      3. Automobile insurance/warranty
      4. Home insurance/renters insurance
      5. Flood insurance
      6. Other insurance/warranty
3. **Educational Services (61-NAICS)**
   1. School notification calls (schools and colleges)
      1. Missed classes
      2. Missed homework
      3. School closure
      4. Other school notification
4. **Health Care and Social Assistance (62-NAICS)**
   1. COVID related calls
      1. COVID case notification
      2. COVID testing updates
      3. COVID awareness
      4. COVID vaccine
      5. Other calls that mention COVID or the pandemic
   2. Healthcare and social assistance
      1. Pharmacy notifications/prescription reminders (CVS, Walgreens, etc.)
      2. Senior care, old age home and residential care
      3. Diagnostic labs
      4. Ambulance services
      5. Dentists, physicians, chiropractors and others
5. **Other Services (except Public Administration) (81-NAICS)**
   1. Political calls
      1. Democrat or republican political robocall
      2. Potential disinformation or voter suppression
      3. Political surveys and voter awareness
      4. Other political robocalls (school elections, local elections, etc.)
   2. Surveys (non-political surveys and non-covid surveys)
   3. Charity and donation calls
      1. Armed forces references
      2. Police references
      3. Fire fighter references
      4. Religious organization/entity references (calls from churches, mosques, temples etc)
      5. Veteran benefits
      6. Other charity or donation calls
   4. Job/employment opportunity leads (not earn-from-home)
   5. Vacation/holiday cruise (scam) calls
   6. Package delivery and shipping company references (FedEx, USPS, UPS, DHL, etc.)
6. **Public Administration (92-NAICS)**
   1. Calls associated with US government entities
      1. Social Security Administration (SSA/SSN)
      2. Social Security Advisor/Disability Advisor
      3. Internal Revenue Service (IRS)
      4. US Treasury
      5. CIA, FBI, NSA
      6. Immigration Department/Department of Homeland Security
      7. Other US Government entity
   2. Calls associated with non-US government entities
      1. Chinese consulate
      2. Canada Revenue Service
      3. Mexico and Mexico specific organization
      4. Other non-US government impersonation
   3. Calls from detention center/prison (calls from inmates)
   4. Jury duty scam
   5. Public interest and awareness calls
      1. Missing person
      2. Lost pet
      3. Adverse weather alert
      4. Other public interest calls
   6. Invitation to join ongoing town-hall conference calls/meetings
7. **Utilities, Transportation, Construction and others (11, 21, 22, 23, 31-33, 42, 44-45, 48-49, 53, 54, 55, 56, 71, 72)**
   1. Utility calls
      1. Electricity (through associated organization: Duke energy, city office etc)
      2. Water
      3. Natural gas connection
      4. Home security and alarm systems
      5. Solar energy systems, wind energy systems
      6. Sewage and water systems
      7. Air conditioning systems and services
      8. Waste disposal and recycling
      9. Cleaning services (carpets, cars, homes etc) and home improvement
      10. Other utility related calls
   2. Farming, agriculture, animal husbandry, forestry
   3. Mining, oil and gas operations (not natural gas lines to homes)
   4. Construction services: Residential and commercial buildings, highways and roads
   5. Manufacturing: Food and beverages, textiles, woodwork, paper and printing, chemicals, vehicle, equipment and others
   6. Retail and wholesale trade: Car dealers, furniture dealers, consumables, clothing, shoes, books, office supplies and others
   7. Transportation: Air, rail, road and water/ferry transportation systems
   8. Press and information publishers: Newspapers, press, TV channels (polls), libraries and others
   9. Hotel, motel, real estate, rental and leasing services (residential buildings, office spaces etc)
   10. Professional and scientific services: Notary, lawyer, accountant, architects and other consultants
   11. Art, entertainment and recreation: dance, music, museums, casinos, theme parks etc
   12. Romance scam/Catfishing
   13. Generic "Can you hear me?" or "Are you there?" lead generation calls
   14. Others calls