Poster: Diving into Robocall Content with SnorCall

Sathvik Prasad
NC State University
snprasad@ncsu.edu

Trevor Dunlap
NC State University
tdunlap@ncsu.edu

Alexander Ross
NC State University
ajross6@ncsu.edu

Bradley Reaves
NC 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. While 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 then 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 analysis of robocall content and trends that can lead to better defenses.**

I. Extended Abstract

Automated phone calls, also called “robocalls”, are a nuisance to every phone user in the United State. Frequent robocalls have made the phone network less trustworthy with people rarely answering important phone calls from unknown numbers. Although robocalls may seem like an annoying distraction to an average phone user, fraudulent robocalls continue to cause significant harm to vulnerable populations in the US. Recent immigrants, senior citizens, students, and Non-English speakers are frequent targets of elaborate scams initiated through robocalls. Such scams often result in significant financial loss, identity theft, or both.

Despite ongoing robocall mitigation efforts by regulators, enforcement agencies, and telecom carriers, fraudulent robocalling operations continue to target phone users by generating millions of robocalls each day. Stakeholders responsible for combating illegal robocalls do not have the necessary human resources or the tools to swiftly analyze large volumes of robocall data and take action against them. Regulators and enforcement agencies manually listen to robocall recordings collected from honeypots or the public. Such a laborious approach to monitoring the robocalling ecosystem is not scalable and often results in delayed enforcement action against the bad actors.

In this work, we present SnorCall, a multi-stage framework for analyzing robocall audio content. By leveraging advancements in semi-supervised machine learning techniques [2], Natural Language Processing (NLP), and robocall campaign analysis [1], SnorCall enables domain experts to swiftly and accurately label robocalling campaigns based by analyzing the audio content of the call. As shown in Figure 1, SnorCall consists of five stages with an additional module to debug the pipeline. We demonstrate the capabilities of SnorCall by analyzing a corpus of over 1.3 million robocalls collected from a telephony honeypot over 23 months.

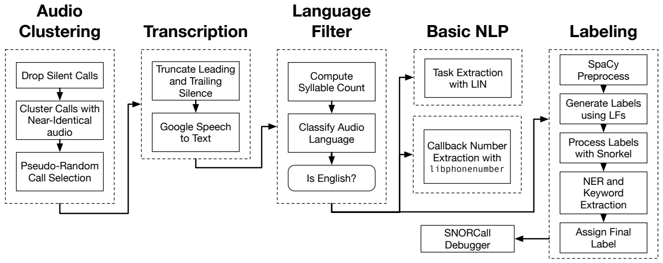


Fig. 1. SnorCall comprises a five-stage pipeline of audio and transcript processing.

By developing five SnorCalls, we study the prevalence of fraudulent robocalling topics (Social Security scams, Tech Support scams, and Financial robocalls), study the evolution of Political robocalls during the 2020 US Presidential Elections, and analyze robocalls designed to target small businesses.

References

[1] Sathvik Prasad, Elijah Bouma-Sims, Athishay Kiran Mylappan, and Bradley Reaves. Who’s calling? characterizing robocalls through audio and metadata analysis. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 397–414. USENIX Association, August 2020.

[2] Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason A. Fries, Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with weak supervision. *VLDB J.*, 29(2-3):709–730, 2020.

We developed a framework to analyze the audio-content of millions of robocalls. We uncovered over 1,000 Social Security campaigns and found that Tech Support scammers try to make about $400 from each scam call.

Diving into Robocall Content with SnorCall

Sathvik Prasad, Trevor Dunlap, Alexander Ross, Bradley Reaves

Introduction

- **Robocalls** or pre-recorded spam calls are a menace for phone users in the United States and have undermined the trustworthiness and utility of legitimate phone calls
- Telecom carriers, regulators, and enforcement agencies are struggling to stop illegal robocalling operations because they don't have the tools to swiftly analyze large volumes of robocall recordings (millions of calls per day)
- We design a system to extract insights from large quantities of robocall recordings using a semi-supervised ML framework called Snorkel and other NLP techniques
- Among many other findings, we present the first-ever estimates of the prevalence of various robocall topics, study the tactics used by government impersonation robocalls, estimate the dollar amounts solicited by Tech Support scammers, and monitor the evolution of political robocalls during the 2020 US Presidential Elections

Methods

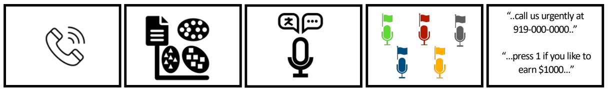
- By operating a honeypot with 6k phone numbers, we recorded more than 1.3 Million robocalls over a 23-month period and uncovered 27K robocalling campaigns**
- Using Snorkel's semi-supervised ML framework, we developed a highly accurate pipeline to swiftly label robocall transcripts with minimal effort and training data
- We reliably extract "callback numbers" tied to robocalling infrastructure, and study deception tactics involving impersonation of government entities, consumer tech companies, well-known e-commerce brands and services

Results

- Social Security scammers tend to operate from office-like infrastructure and their operations were disrupted due to COVID lockdowns. They falsely associate themselves with other federal agencies (FBI, DEA, US Treasury) and target disabled people, senior citizens, and recent immigrants
- Thousands of robocalls misrepresented political events to potentially steal personal information by falsely advertising "..student loan forgiveness program by the Biden Administration.." which was non-existent in August 2021
- Tech Support scam calls try to defraud victims of $400 on average by impersonating Amazon Prime and Apple iCloud customer support agents. They are moving away from well-known Microsoft/Windows tech support tactics
- Fraudulent financial robocalls target distressed taxpayers in the United States. They impersonate well-known banks, credit card companies, cryptocurrency wallets, and the IRS
- Robocalls frequently use "callback numbers" to engage with their targets. These numbers are short-lived and are shared between seemingly unrelated campaigns, indicating potential infrastructure reuse among robocalling operations
- We uncovered thousands of campaigns targeting Mandarin and Spanish-speaking populations. They impersonate government embassies and banks in the US

Takeaways

- Our framework empowers regulators, investigators, and carriers to proactively uncover malicious robocalls and prioritize the takedown of egregious robocalling operations



"..call us urgently at 919-000-0000.."

"...press 1 if you like to earn $1000..."

Honeypot Data Collection

~6,000 phone numbers

23 months of data

1.3 Million robocalls

Robocall Campaign Detection

26,800 Campaigns

370K long-duration robocall recordings

Transcription and Language Detection

25,200 English Campaigns

1,600 Non-English Campaigns

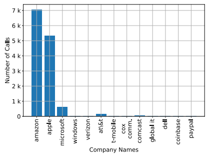
Campaign Labeling using Snorkel and NLP

- 3.6% SSN Scams
- 3.7% Tech Support
- 5.2% Political
- 25% Financial
- 10% Business Listing

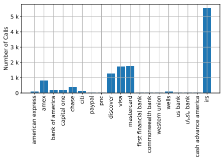
Extracting Operational Characteristics

~45% calls contain callback numbers

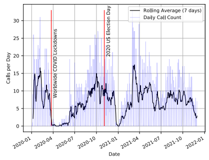
Unrelated campaigns share callback numbers



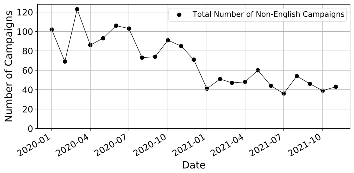
Amazon tech support scams are the most popular type of tech support campaign



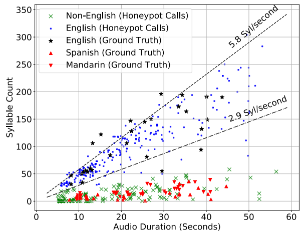
The IRS, credit card companies and banks are common targets of impersonation scams



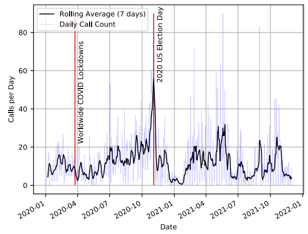
Social Security scam calls were severely impacted by COVID lockdowns



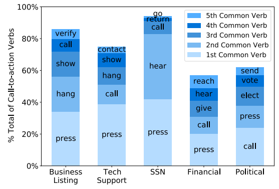
Non-English robocalls (Spanish, Mandarin etc) are a significant threat to vulnerable populations



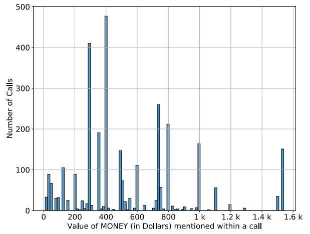
Our language-detection heuristic accurately classifies English and Non-English Robocalls



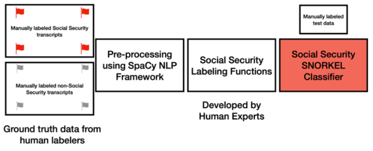
Political robocalls peaked during the 2020 US Presidential Elections (3rd Nov 2020)



Instructing call recipients to press a digit is the most common call-to-action



Distribution of Money (in US Dollars) within Tech Support robocall with a median of $400


Manually labeled Social Security transcripts
Manually labeled non-Social Security transcripts
Pre-processing using SpaCy NLP Framework
Social Security Labeling Functions
Social Security SNORKEL Classifier
Manually labeled test data
Developed by Human Experts
Ground truth data from human labelers

Iterative trailing/development pipeline to develop a Social Security SnorCall

** Details about our honeypot and the language-agnostic robocall campaign identification technique based on audio-similarity is described in our previously published 2020 USENIX Security paper: "**Who's Calling? Characterizing Robocalls through Audio and Metadata Analysis**"

NC STATE UNIVERSITY
NSF
WSPR