Abstract—Unsolicited bulk telephone calls — termed “robo-
calls” — nearly outnumber legitimate calls, overwhelming tele-
phone 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 signifi-
cant financial loss, identity theft, or both.

Despite ongoing robocall mitigation efforts by regulators,
enforcement agencies, and telecom carriers, fraudulent robo-
calling 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 ap-
proach 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 advance-
ments in semi-supervised machine learning techniques [2],
Natural Language Processing (NLP), and robocall campaign
analysis [1]. SnorCall enables domain experts to swiftly and

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

Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with
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 law 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 Snorrel 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 Snorrel’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.
- 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 Snorrel’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.

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