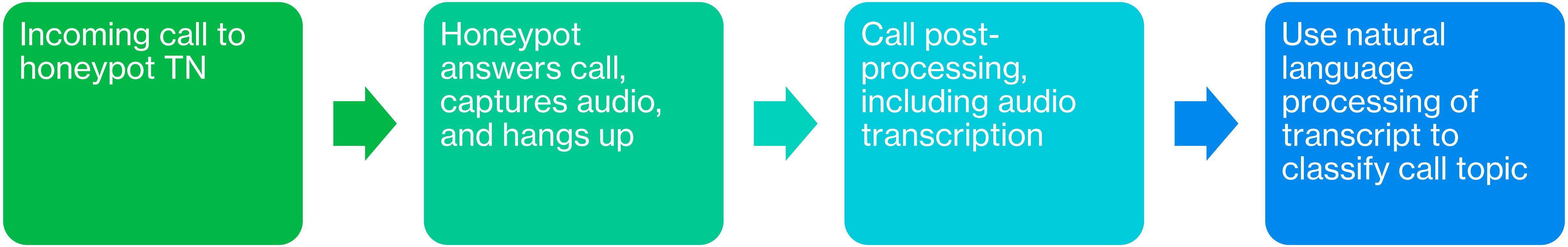
Using natural language processing to label robocalls

Kate Kutchko
Verizon Business Group

Honeypot introduction

Verizon’s telephony honeypot

- **Thousands of honeypot TNs, mostly in wireless ranges**
- **Coverage in nearly every US NPA**



Incoming call to honeypot TN

Honeypot answers call, captures audio, and hangs up

Call post-processing, including audio transcription

Use natural language processing of transcript to classify call topic

Goal of natural language processing (NLP) is to identify and label robocall campaigns

Transcript	Campaign
Hello, this is Margaret with Economic Impact and Alternative Loan Assistance, how are you doing today? Okay I only need a moment of your time. I am actually working on getting some uh important information over to you about the new economic impact debt elimination and alternative loan program. Uh, the economic impact Loan program has –	loan: economic impact
My name is Olivia. I am a senior settlement officer from our tax division. How are you doing today? Good to hear that. We have noticed you have a back tax debt that needs to be settled before the end of the month.	tax debt/IRS: senior settlement officer
Hello this is a National Police and Trooper Association. We're calling everyone to let them know the –	fundraising: natl police and trooper association
Hi this is Crystal with riseup debt solutions on a recorded line. How are you doing today?	debt: rise up debt solutions
Hey this is Eva with Your Verified Now. Our records show your Google Business listing is not properly verified with Google. This can cause customers searching for your services to not be able to find –	business listing: google business listing

Natural language processing

Natural language processing using TF-IDF

$$TF - IDF(term, doc) = TF(term, doc) * IDF\ (term)$$

Term frequency: Count of appearances of a word in a transcript (document)

Document frequency: Number of honeypot transcripts (documents) that word appears in

Common words like “please,” “call,” “number” have a **high document frequency**, so they are **less important** for transcript classification

Rare words like “payday,” “security,” “cruise” have a **low document frequency**, so they are **more relevant** for transcript classification

With TF-IDF (term frequency – inverse document frequency), we create a vector for each honeypot transcript based on the words in each transcript.

The cosine distance measures the differences between two transcripts

Identical robocalls may have differences in transcription because of:

- **Audio cutting off at a different point**
- **Errors/inconsistencies in transcription**
- **Repetition of parts of the robocall recording**

In addition, robocalling campaigns often make slight changes to their scripts, such as day of the week or caller name

Cosine distance ranges from 0 to 1

0.0

1.0

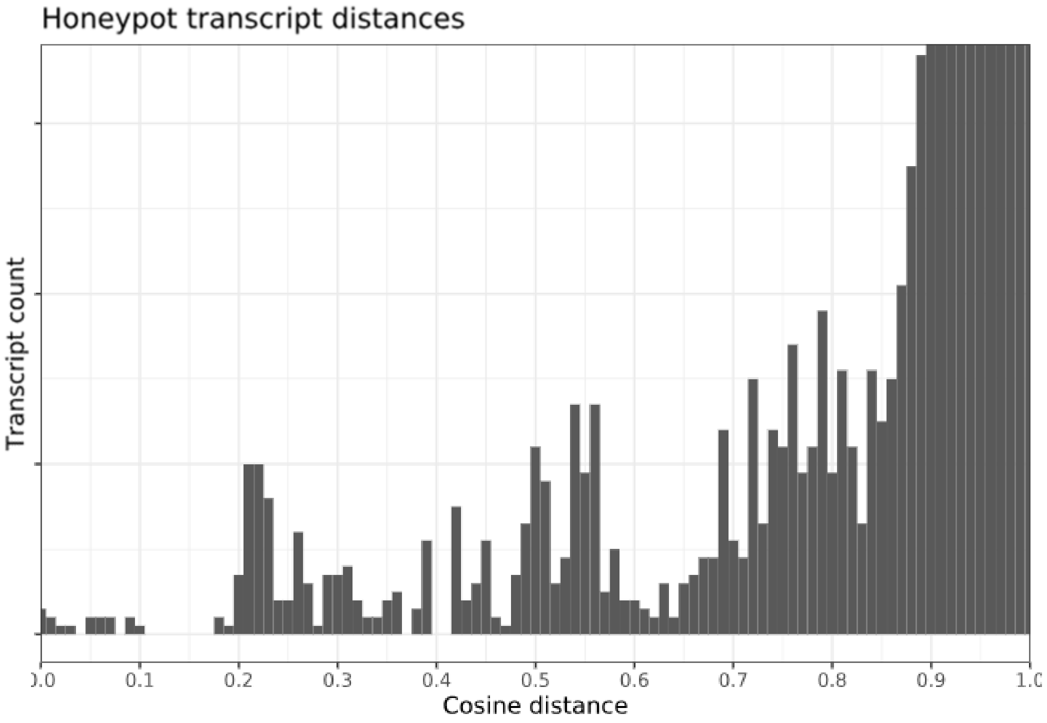
Identical transcripts

No words in common

Examples of transcription matches

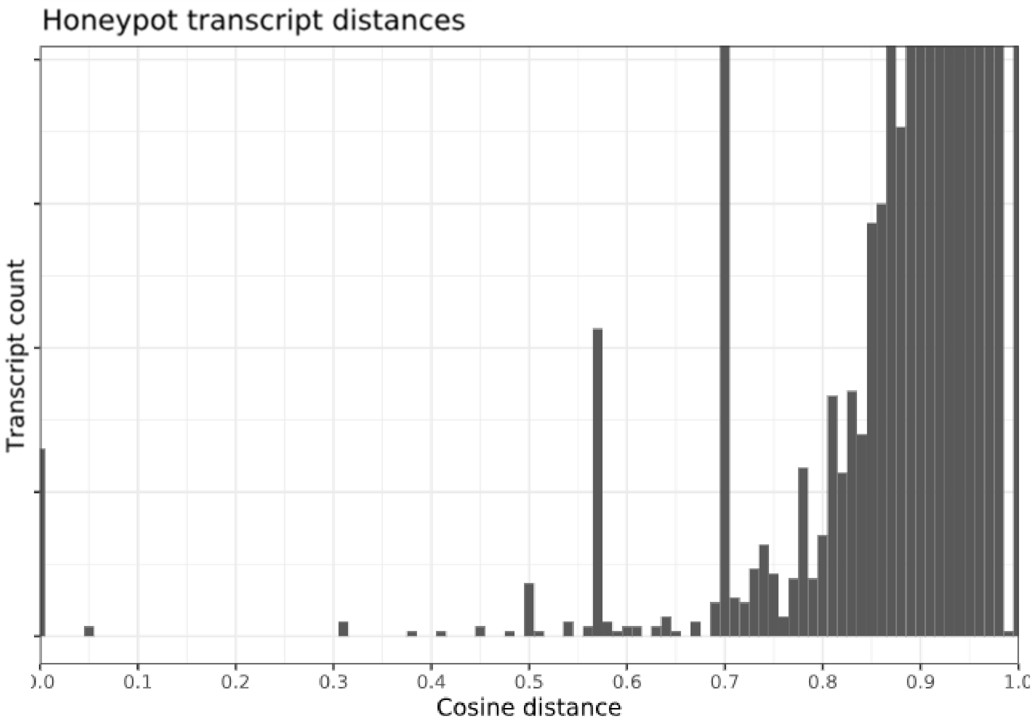
Example: Economic Impact Relief Center

Cosine distance	Transcript
0.00 (reference transcript)	This is an important update from the Economic Impact Relief Center in regards to your personal loan application. Hello this is an important update from the Economic Impact Relief Center regarding your personal loan application. Our contact number is 844-XXX-XXXX. We are reaching out to inform you that our underwriting department –
0.21	Please do not hang up. This is an important message from the Economic Impact Relief Center regarding your application. This is an important update from the Economic Impact Relief Center regarding your economic impact relief application. Phone number 833-XXX-XXXX special enrollment period for the economic impact relief –
0.42	Please do not hang up this is an important message from the Economic Impact Release Center regarding your application. Phone number 833-XXX-XXXX special enrollment period for the economic impact relief program is coming to an end and we're missing information from you. Please press 2 to speak to a member of our team or 9 to be placed –
0.51	Hello this is Margaret with economic impact and alternative loan assistance how are you doing today? Okay I only need a moment of your time. I am actually working on getting some important information over to you about the new economic impact loan elimination and alternative loan program the –



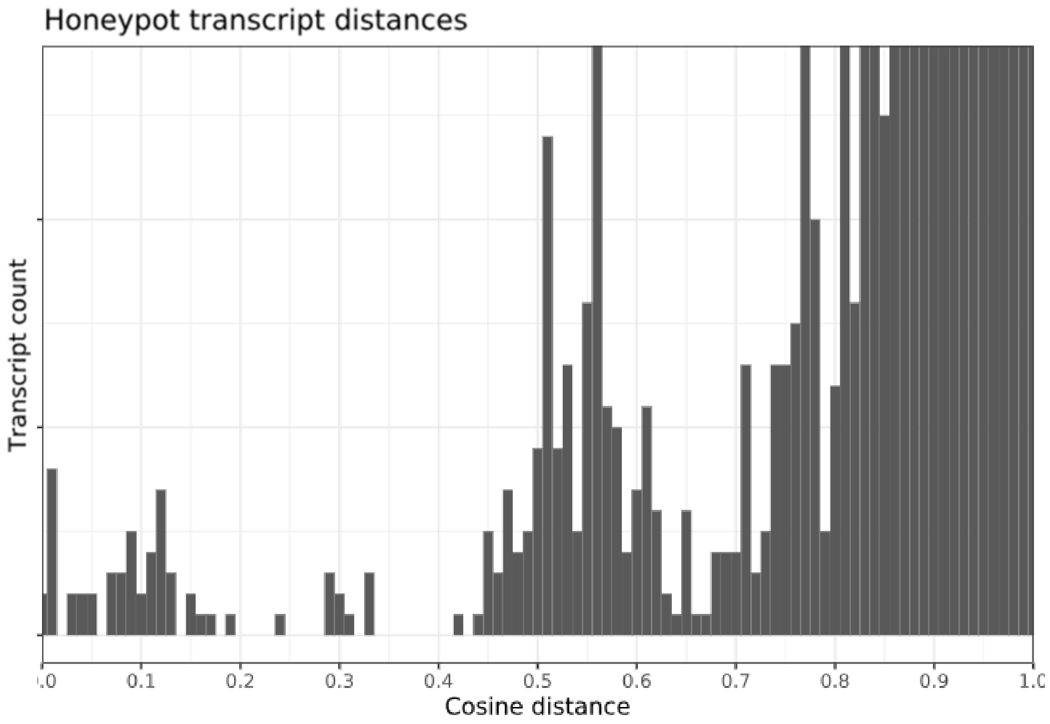
Example: Home Security Promotions

Cosine distance	Transcript
0.00 (reference transcript)	Hi my name is Cassidy with home security promotions. How are you today?
0.05	Hi my name is Cassidy with home security promotions. How –
0.31	Hello, hi my name is Cassidy I was home security –
0.50	Hi my name is Dave with General Electric home security. How are you doing today?
0.57	Hello my name is... How are you doing today?



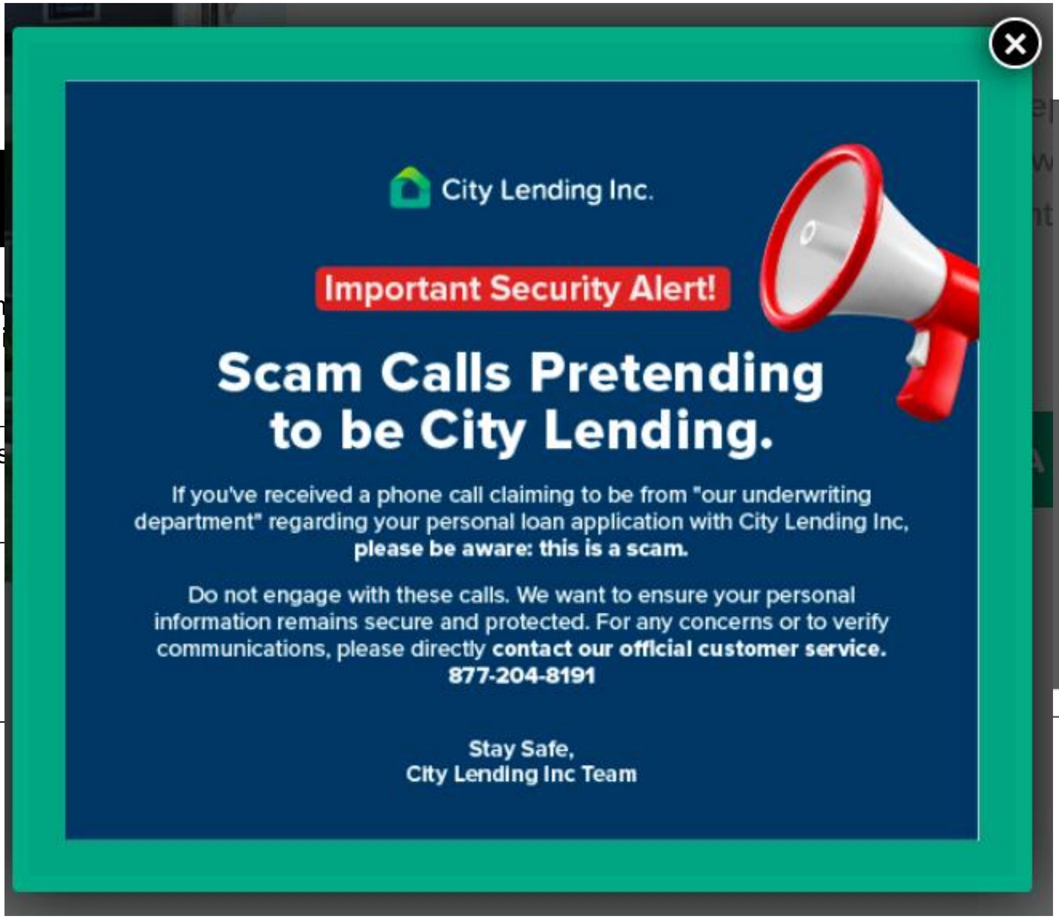
Example: City Lending

Cosine distance	Transcript
0.00 (reference transcript)	This is an important update from City Lending in regards to your personal loan application. Phone number 844 – Hello this is an important update from City Lending regarding your personal loan application. Our contact number is 844-XXX-XXXX. We are reaching out to inform you that our underwriting department is currently missing certain items required –
0.16	This is an important update from City Lending in regards to your – Hello this is an important update from City Lending regarding your personal loan application. Our contact number is 844-XXX-XXXX.
0.45	Hello this is a message from First Premier Lending. We're calling regarding a critical update on your loan application. Hello this is Ciara with lending agent ID 3463 reaching out regarding your loan application. I see here we're missing a couple pieces of information to complete the application but so far –



Example: City Lending

Cosine distance	Transcript
0.00 (reference transcript)	This is an important update from City Lending in regards to your personal loan application. Phone number 844 – Hello this is an important update fron City Lending regarding your personal loan application. Our contact number i 844-XXX-XXXX. We are reaching out to inform you that our underwriting department is currently missing certain items required –
0.16	This is an important update from City Lending in regards to your – Hello this is an important update from City Lending regarding your personal loan application. Our contact number is 844-XXX-XXXX.
0.45	Hello this is a message from First Premier Lending. We're calling regarding a critical update on your loan application. Hello this is Ciara with lending agent ID 3463 reaching out regarding your loan application. I see here we're missing a couple pieces of information to complete the application but so far –



City Lending Inc.

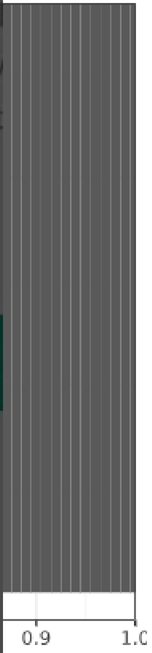
Important Security Alert!

Scam Calls Pretending to be City Lending.

If you've received a phone call claiming to be from "our underwriting department" regarding your personal loan application with City Lending Inc, **please be aware: this is a scam.**

Do not engage with these calls. We want to ensure your personal information remains secure and protected. For any concerns or to verify communications, please directly **contact our official customer service. 877-204-8191**

Stay Safe,
City Lending Inc Team



Process to add a new campaign (City Lending)

1. **Select the reference transcript for that campaign**

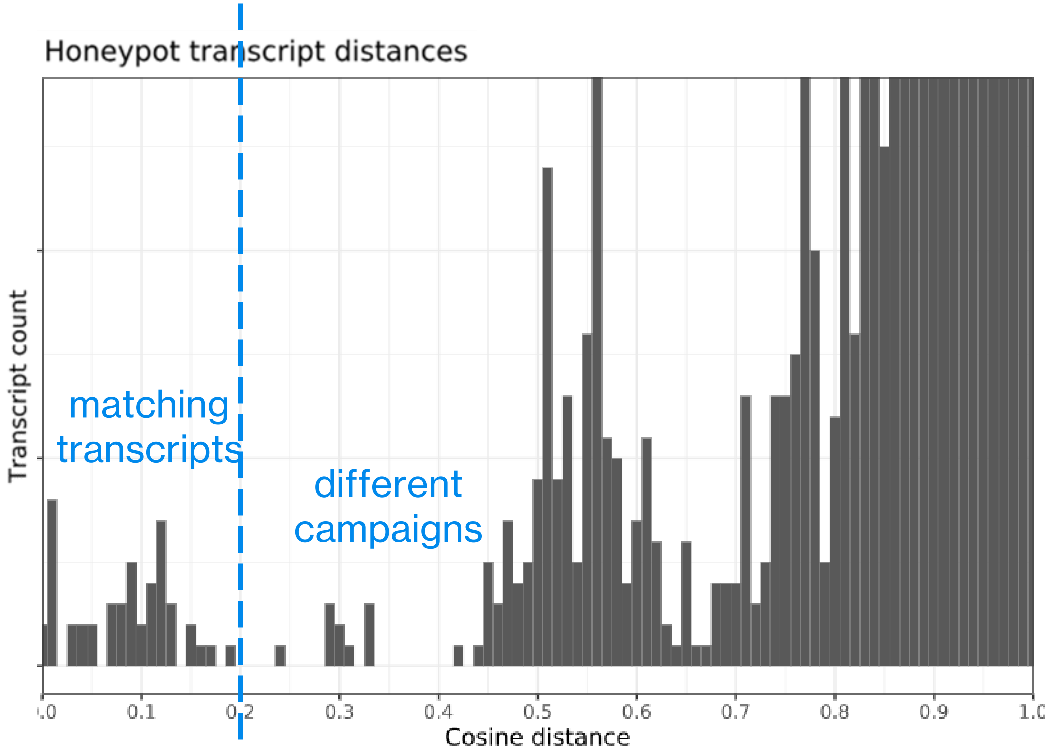
   “This is an important update from City lending in regards to your personal loan application...”

2. **Compare the cosine distance of that reference transcript with other call transcripts to get a distribution**

3. **Identify the upper cosine distance limit (blue dotted line)**

   If a transcript has a cosine distance less than the limit (compared with the reference campaign), it belongs to that campaign

   If a transcript has a cosine distance greater than the limit, it does not belong to that campaign *(but it may belong to a related campaign)*



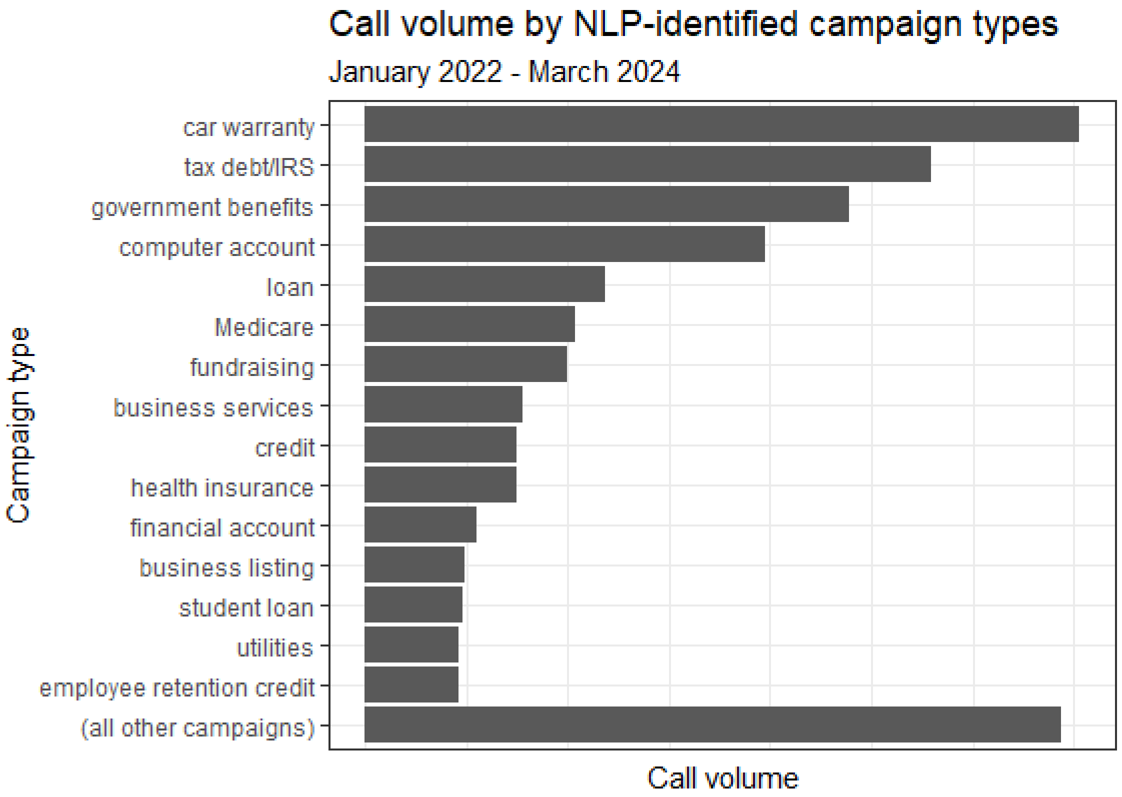
Campaign trends

This approach can automatically identify good robocalling candidates for traceback

By choosing conservative cosine distance limits, we minimize false positives in campaign labels

We label incoming calls automatically and flag as candidates for traceback through the Industry Traceback Group

Since January 2022, we have identified thousands of calls belonging to over 400 robocalling campaigns in our honeypot through these methods


Call volume by NLP-identified campaign types

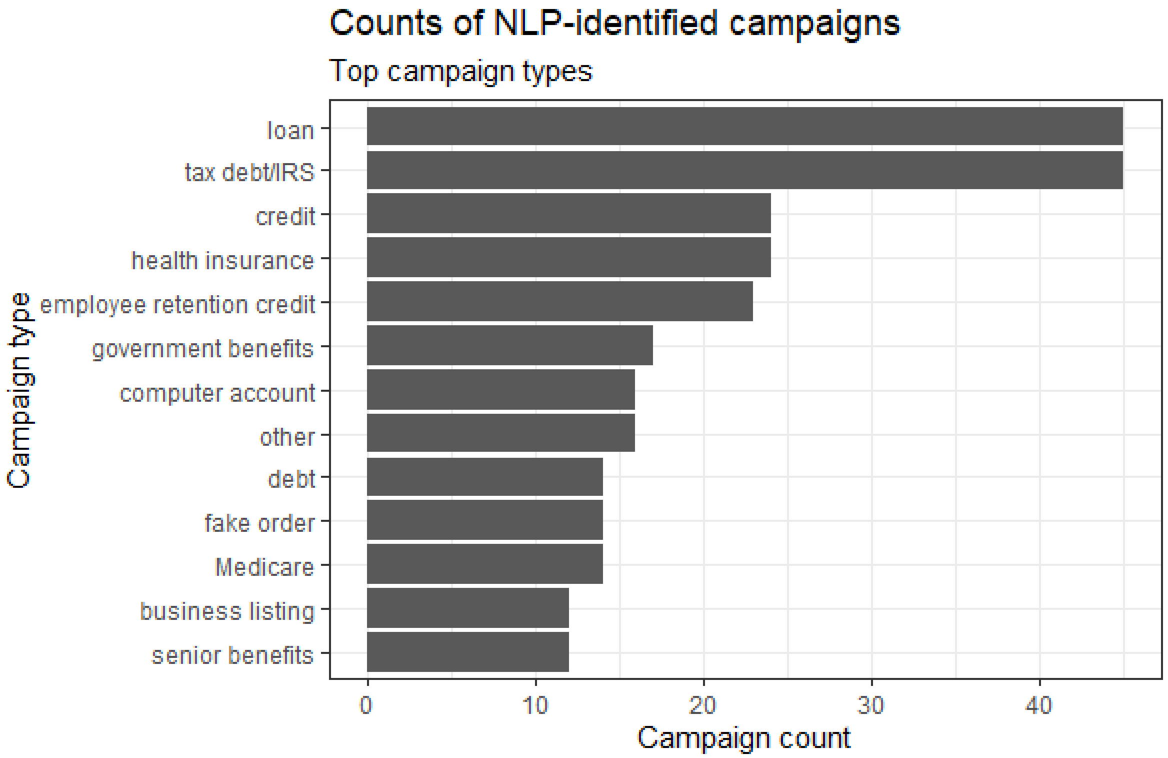
January 2022 - March 2024

Campaign counts by campaign type

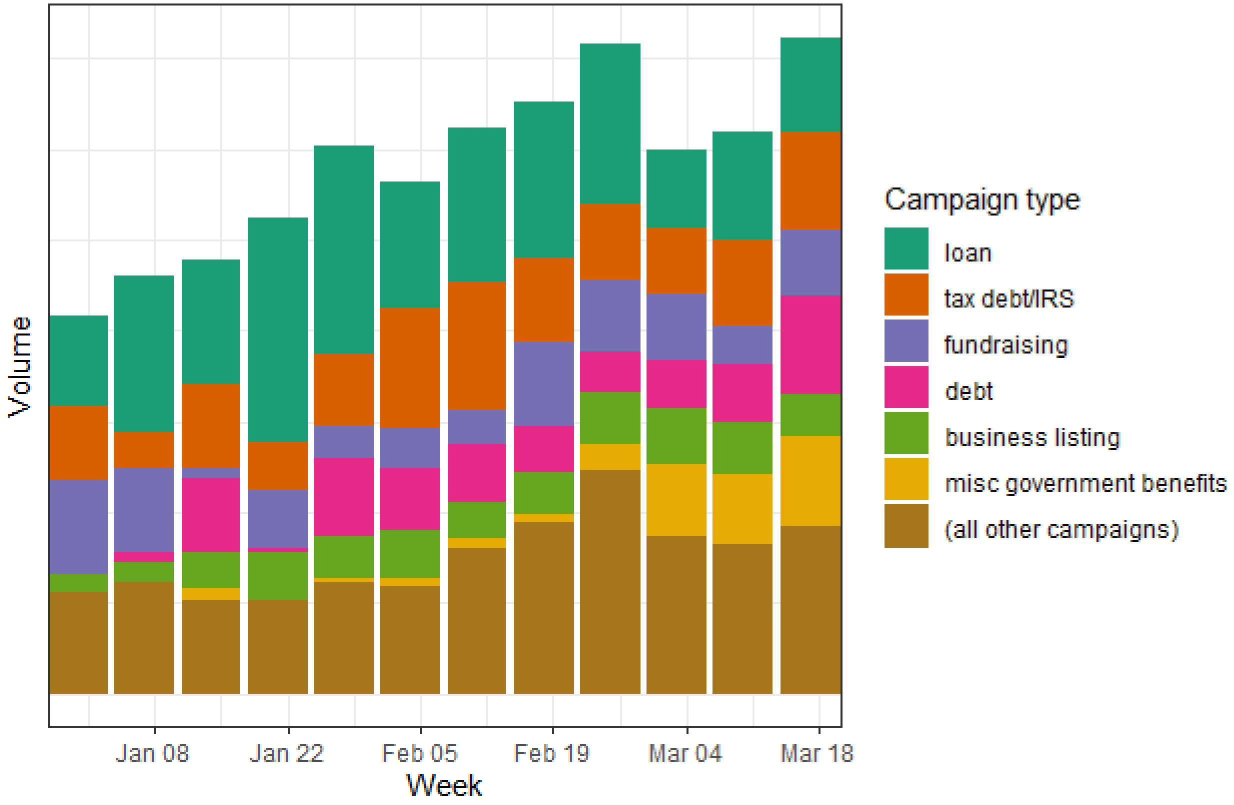
We see the most variety in campaigns for “loan” and “tax debt/IRS” campaigns

Recent uptick in financial-related campaigns

Lots of variation in entity name and robocalling transcript



We are currently seeing a wide and distributed variety of robocalling campaigns to the honeypot



Weekly volume by campaign type

Early 2024

Campaign type	Distinct campaigns seen in honeypot (early 2024)
loan	22
tax debt/IRS	18
fundraising	6
debt	10
business listing	7
misc government benefits	6
(all other campaigns)	70

Takeaways

Takeaways

We can use natural language processing techniques such as TF-IDF as effective tools for robocalling transcript classification, in particular for calls to our voice honeypot.

By choosing conservative cosine distance limits, we can be confident that labeled transcripts belong to the correct reference campaign.

We are seeing a wide variety of robocalls in our honeypot in recent months, with a specific focus on robocalls relating to debt and personal loans.

Thank you!