Imperial College London

Department of Earth Science and Engineering

MSc in Applied Computational Science and Engineering

Independent Research Project

Pattern Mining Phase Summary

Unlocking Global Capital: AI-Powered Prediction, Access, and Verification of Investor Decision Makers

by

Daniel Bowman

Email: daniel.bowman24@imperial.ac.uk

GitHub username: acse-db1724

Repository: ese-ada-lovelace-2024/irp-db1724

Supervisors:

Antony Sommerfeld

Dr. Yves Plancherel

June 2024

Table of Contents

List of Tables

List of Figures

Abstract

Automating investor contact maintenance is critical for scalable, reliable capital raising in a sector plagued by high turnover and data decay. Existing commercial services like Hunter and ZoomInfo rely on paid lookups and often take days to return results. In contrast, our novel pipeline seamlessly integrates three components—a comprehensive offline template miner that uncovers every common formatting skeleton, a lightweight real-time classifier that instantly predicts the correct template for any new name–domain pair, and on-the-fly third-party deliverability scoring—into one end-to-end system. Performance will be evaluated on held-out contacts, reporting template coverage, prediction accuracy, API precision/recall, and sub-50 ms query latency. This hybrid approach not only outperforms standalone pattern-mining or machine-learning methods but also undercuts costly data-provider fees, democratizing access to fresh, validated investor email information.

1. Introduction

This phase of the project focuses on mining syntactic patterns from AIP's dataset of LP investor email addresses. The insights generated here will directly inform the feature engineering process for a downstream LightGBM-based email prediction model. This model aims to democratize access to verified investor contact data across the financial services industry.

The approach blends rule-based logic with data-driven discovery. By tokenizing local-part email structures (e.g. j.smith, john.smith), we generate a compact yet expressive sequence representation. These sequences are then processed using a sequential pattern miner (TRuleGrowth) to extract frequent structural rules. The resulting mined templates reflect dominant naming conventions used across firms and regions—providing both interpretable insights and high-value features for learning-based prediction.

2. Email Tokenisation and Encoding

To tokenise email local parts, the investor was split apart using the python 'NameParser' library which decomposes the name into 'first', 'middle' and 'last' parts. The local part is then scanned for either the full or initial of the name part and a token is then derived to be stored in a sequential list.

In order to get maximum coverage, constant tokens and name variants were introduced to capture cultural naming conventions. For example, surname particles (e.g. van der, de la, etc.) featured heavily in the dataset largely from European investors. A list of common particles were included in the 'EmailTemplateEncoder' class and each of these particles were used to form a surname particle variant. This particle would then result in its own unique token if matched in the email local part. A similar thing was done for 'Nicknames', where common abbreviations or nick names (e.g. 'William' to 'Bill') are included in its own token variant using a lookup hash map.

Accent normalisation was also done to maximise results. This was done in two ways, NFKD normalisation and mapping Germanic characters to their Latin equivalent (e.g. ‘ä’ to ‘ae’), both of these variants formed their own token.

Unknown sequences were marked ‘UNK’ and discarded from the final list. 6,683 unknown sequences were found from a total 81,000, these largely stem from middle names missing from the investor name field but appearing the local email part. These sequences are captured within the ‘EmailTemplateEncoder’ class for analysis.

The mined sequences are then put into a set so a unique list can be derived allowing us to see some of the template diversity per firm.

firm	num_templates	top_template	top_template_share
0704 capital	1	('f_0', 'last_original_0')	1
1 north wealth services	2	('f_0', 'last_original_0')	0.8
10 branch	1	('first_original_0',)	1
10 east	1	('f_0', 'last_original_0')	1
1010 capital	1	('first_original_0',)	0.875
1199 seiu regional pension fund	1	('first_original_0', '.', 'last_original_0')	1
1199seiu national benefit fund	1	('first_original_0', '.', 'last_original_0')	0.8
13th floor capital	1	('f_0', 'last_original_0')	1
1492 capital management	1	('f_0', 'last_original_0')	1
1607 capital partners	1	('f_0', 'last_original_0')	1
1650 wealth management	1	('first_original_0', 'last_original_0')	1
1776 wealth	1	('first_original_0',)	1
1788 capital	1	('first_original_0', '.', 'last_original_0')	1
180 degree capital	2	('first_original_0',)	0.5
1832 asset management u.s	2	('first_original_0', '.', 'last_original_0')	0.833333
1834 investment advisors	2	('f_0', 'last_original_0')	0.5
1858 wealth management	1	('first_original_0',)	1
1875 finance	2	('f_0', 'last_original_0')	0.857143

1888 management	2	('first_original_0', '.', 'last_original_0')	0.666667
1900 wealth	2	('first_original_0',)	0.846154

Table 1: Template Diversity per Firm

At a first glance it would seem that most firms use just one template type, highlighting a potential lack of template diversity per firm, which makes the job of predicting emails that much easier. Further analysis highlights that the majority of the 404 mined unique sequences only cover 0-5% of the dataset, with the vast majority of the dataset being covered by just a few template types. This indicates a Pareto or long tail distribution in the template coverage.

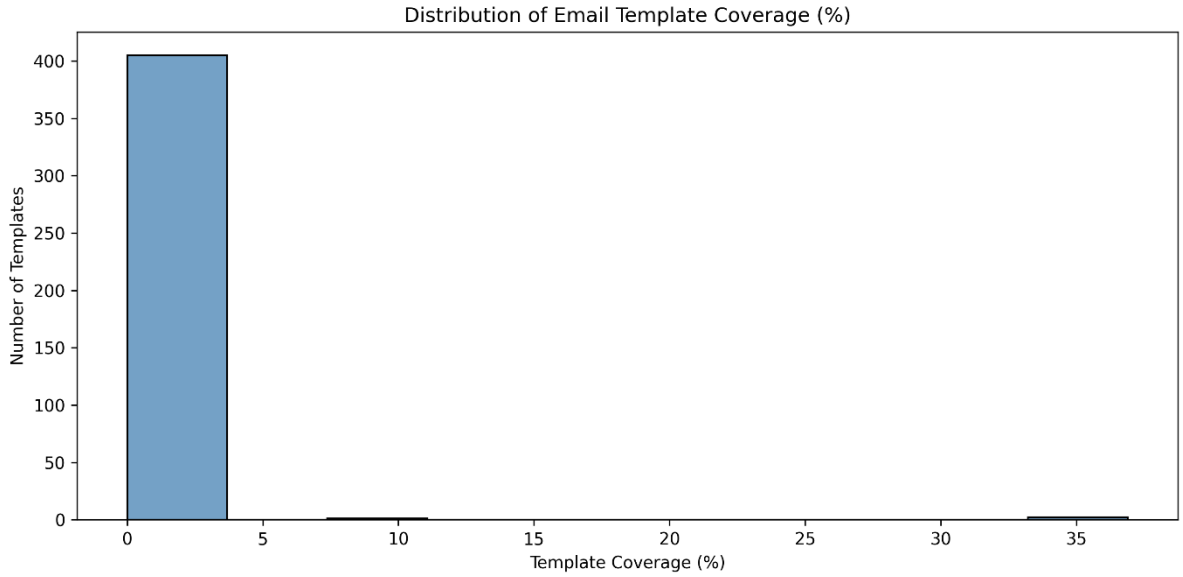

Figure 1: Distribution of Template Coverage

3. Template Rule Mining

To better understand dominant email structures across firms, we applied sequential rule mining using the TRuleGrowth algorithm from the SPMF Library. This library was created by Philippe Fournier-Viger who wrote the original paper that inspired this use of miner during the literature review. The actual algorithm was implemented in Java and the Python library ‘Spmf’ provides a lightweight wrapper for it.

This algorithm was chosen as it captures token order, supports mining rules with support and confidence thresholds and is efficient especially with our categorizable and sequential data.

3.1. Dominant Rules

Mined rules are expressed with antecedent and consequent. The antecedent is at times referred to as the LHS whereas the consequent is called the RHS and these two come together to form a item sequence where the LHS appears prior to the RHS in the rule. The support (amount of cases the rule was mined from) and confidence (probability that the rule appeared in those cases) then quantify the rule.

Some of the most frequent rules feature the first name (indexed at 0 in cases of double barrelled or multiple first names) concatenated with a '.' separator appear before the last name in 97% of relevant cases. This indicates that names like John Smith typically map to the email local "john.smith" in 97% of relevant cases. There are also rules that utilise initials of both the first and last name to form the email local part, as well as rules referring to middle names and multiple last names that found their way past the confidence threshold.

a) Impact on Feature Engineering

Ultimately, our feature matrix to the prediction engine will have a list of candidate sequences – derived from the unique templates mined from the dataset – which the model will assign a probability to. These mined rules allow us to flag candidate sequences with template confidence, support, and other metadata regarding how the mined rules support the template.

4. Extended Exploratory Data Analysis (EDA)

Further EDA was done on our mined template to see how diverse our dataset is.

4.1. Template Diversity Across Firms

Analysis highlights a lack of template diversity in firms. With the majority of firms using just one template.

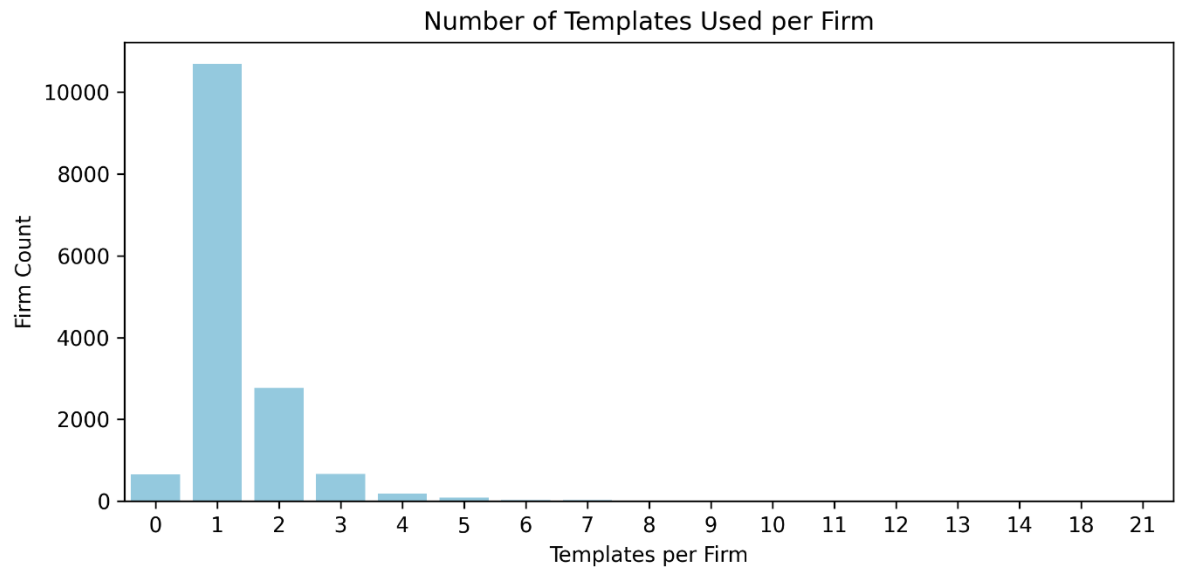

Figure 2: Number of Templates per Firm

Once again this is distributed with a long tail leaving us with some firms having as many as 21 email templates.

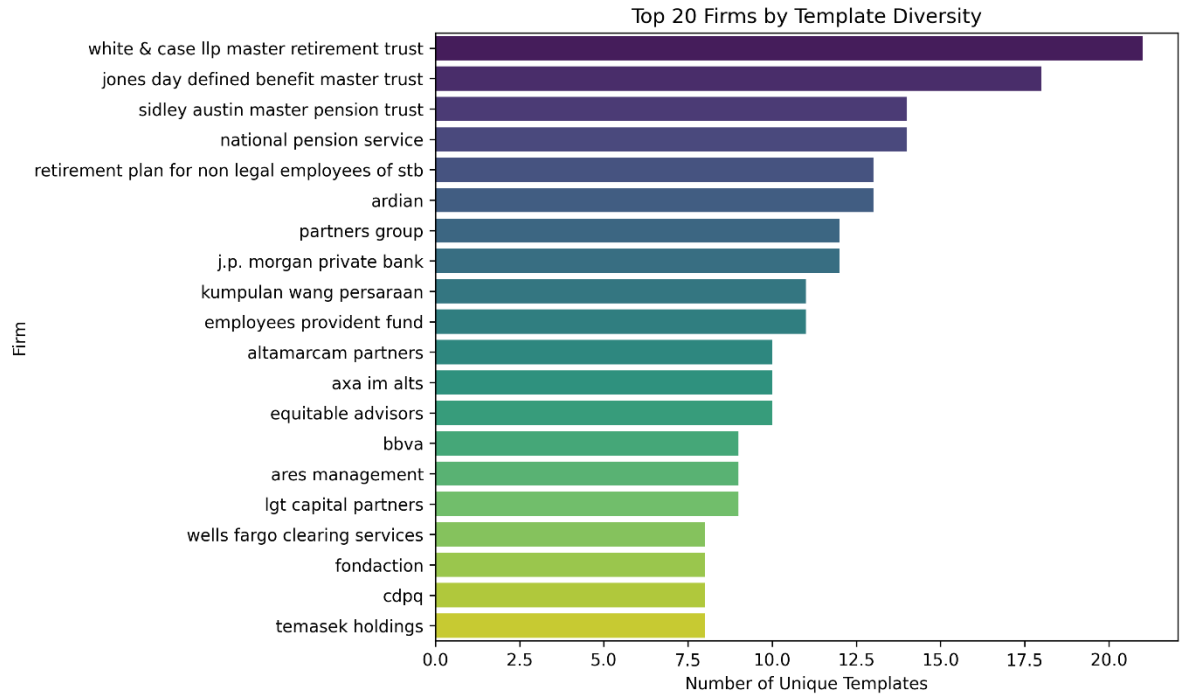

Figure 3: Top 20 Firms by Template Diversity

Despite these outliers, the statistics tell us that majority of firms use just one template structure.

	num_templates
count	15122
mean	1.321716704
std	0.878952738
min	0
25%	1
50%	1
75%	2
max	21

Table 2: Number of Template Statistics

4.2. Name Structure Complexity

Middle names, multiple names (for middle, first and last names) and other name structures are flagged to enrich our feature matrix. We also used these flags to investigate template usage amongst firms.

The instinct was that template diversity was driven by name complexity, i.e. nickname usage or double barrelled names being parsed into the email local part.

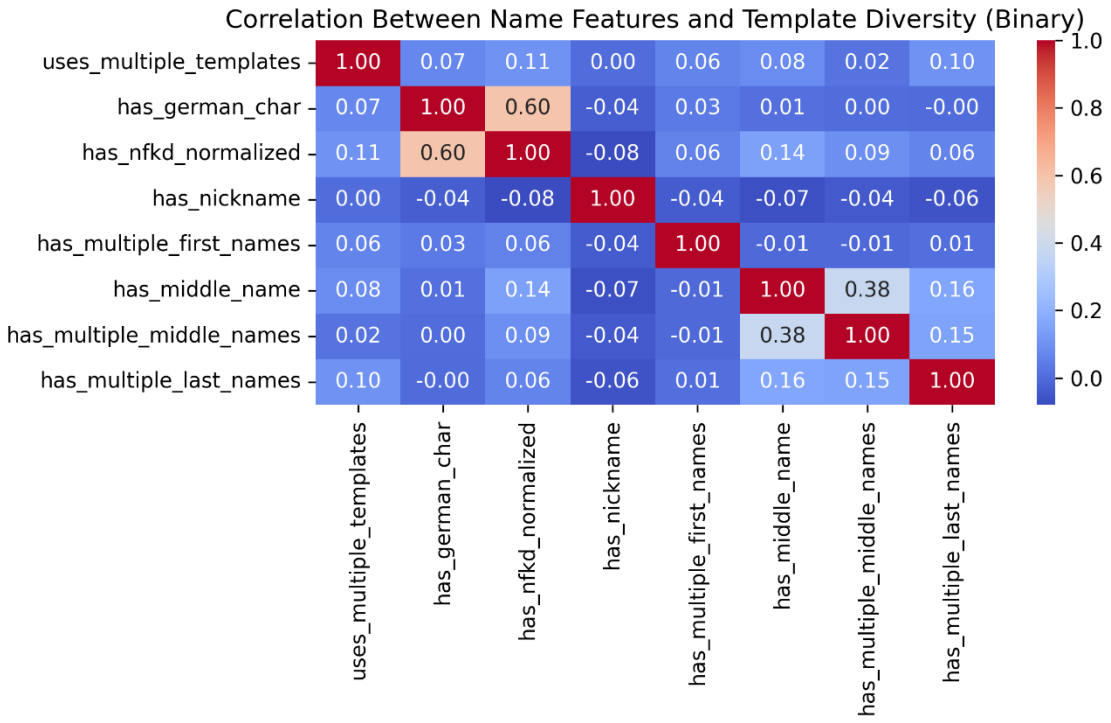

Figure 4: Name Complexity and Template Diversity Correlation

Despite this, there are no overwhelmingly strong correlations between name complexity and template diversity. There are a few signals leaning towards names requiring NFKD normalisation and investors that have multiple last names (most of these cases are down to cultural naming conventions) which supports the choice to include these flags into the feature matrix.

4.3. Firm Size and Template Diversity

Firm size was also explored, the suspicion was that complex templates were chosen to account for multiple investors with similar names (i.e. the same initials of same first and last name).

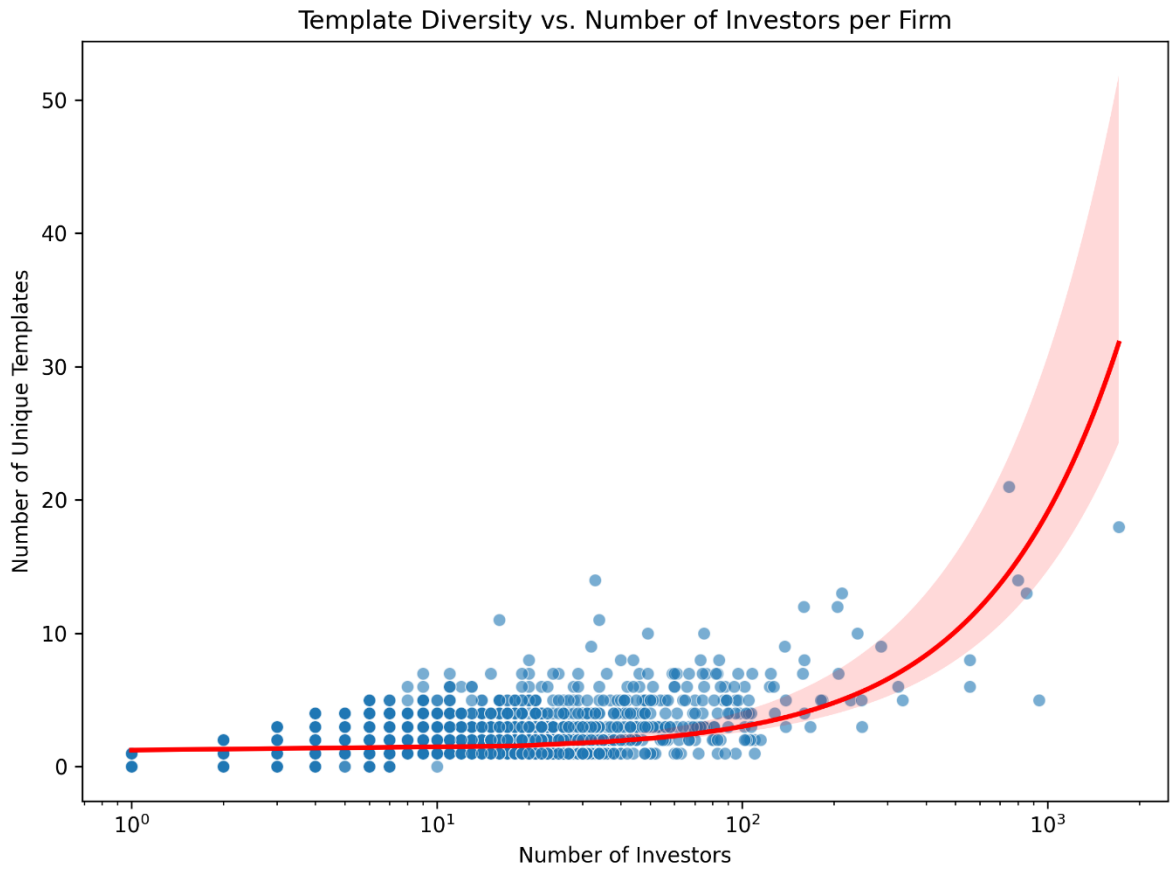

Figure 5: Template Diversity vs. Number of Investors

Analysis revealed that there is a clear trend towards increased template diversity and firm size, especially in the extreme cases. However, generally firms stick to just one or two templates unless they need to divert. Heuristically we can infer that a mixture of name complexity and firm size is responsible for template diversity.

4.4. Field Level Template Analysis

Role and country distributions were analysed for template entropy.

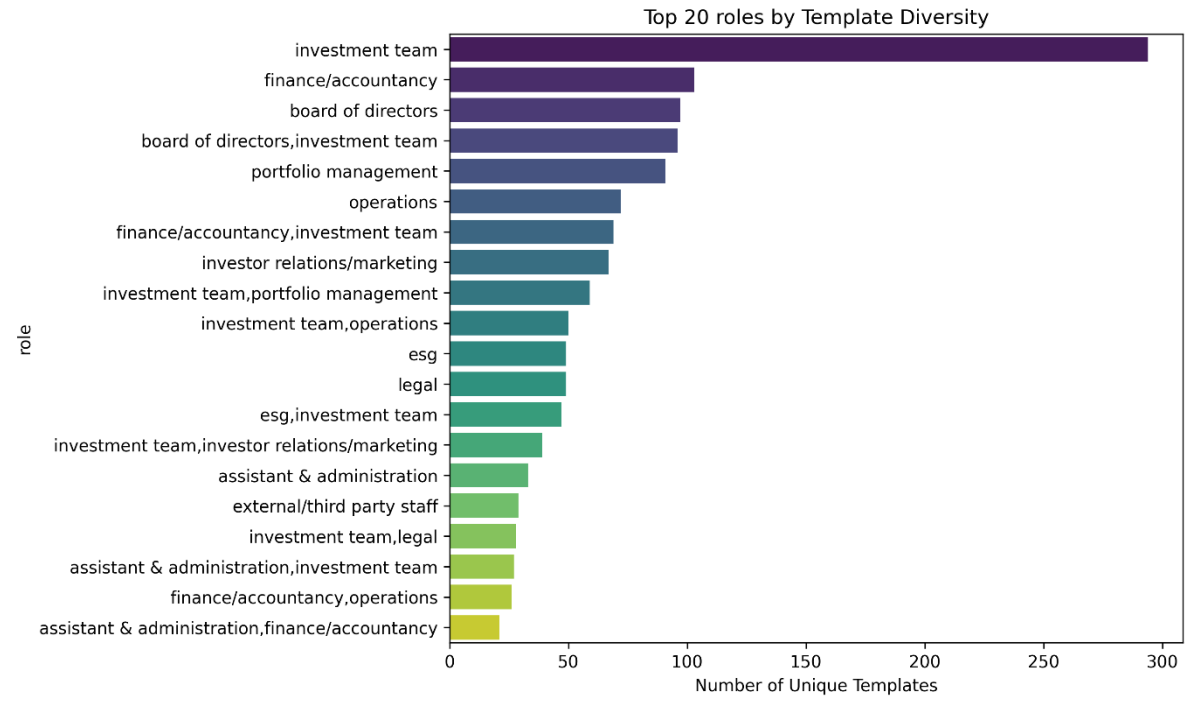

Figure 6: Top 20 Roles by Template Diversity

Clearly members of an investment team encounter high template diversity, probably down to it being a role that features a lot of investors (as opposed to legal or executive teams).

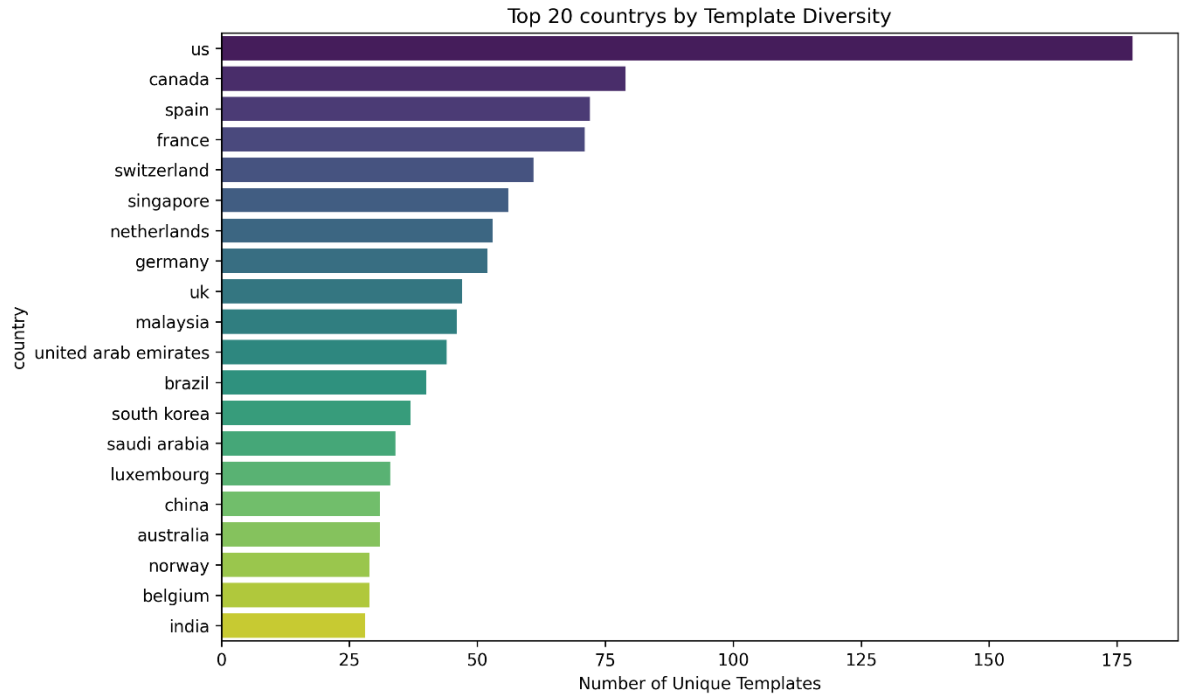

Figure 7: Top 20 Countries by Template Diversity

Similar trends by country, can largely be attributed to the US being a hub for global investment.

5. Domain Patterns and Coverage

In addition to email local part structure, email domain mappings to firms names were also investigated. This was done to inform our approach towards domain prediction for unseen firms. For context, AIP have hinted that the list of firms provided is exhaustive and we don't expect to see unseen firms at inference (and if we do we assume that predictions are entirely unreliable).

Using a similar firm name encoder as was used during email template discovery, firm names were tokenised and passed into the same TRuleGrowth sequential miner. Firms were tokenised into meaningful substrings (words, initials, suffixes, etc) to capture structures in the email domain.

Despite efforts taken to maximise coverage, almost half of the domains were not tokenised due to unintuitive domain structures or shared domain structures were subsidiaries use domain that relates to their parent.

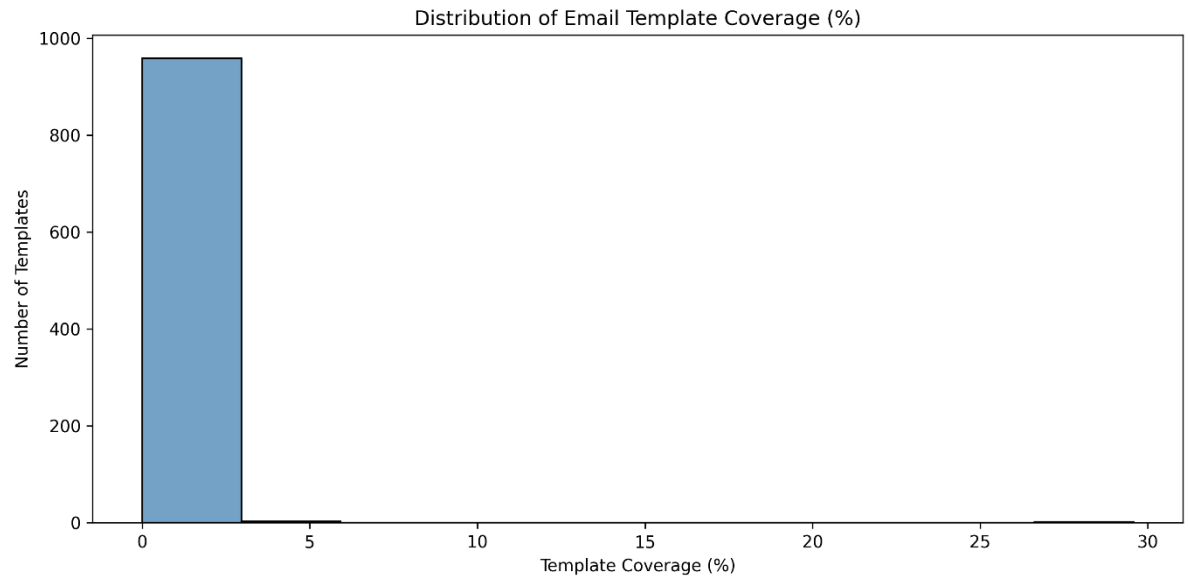

Figure 8: Distribution of Domain Templates

Coverage trends the same way email templates, however it is important to keep in mind that there are almost half the mined templates for domains as opposed to local parts. Also there are double the amount of unique templates mined from email domains.

Domain template missingness was investigated against firm structure flags, highlighting a clear trend that shows complex corporate structures make predicting domain mappings difficult. As a result, these findings are preserved but do not provide a feasible path towards domain prediction and are effectively shelved.

5.1. Domain-Website Similarity

To further understand domain reliability, we compared each firm's email domain to the root of the listed website using Python's 'SequenceMatcher' to compute a similarity score for each pair.

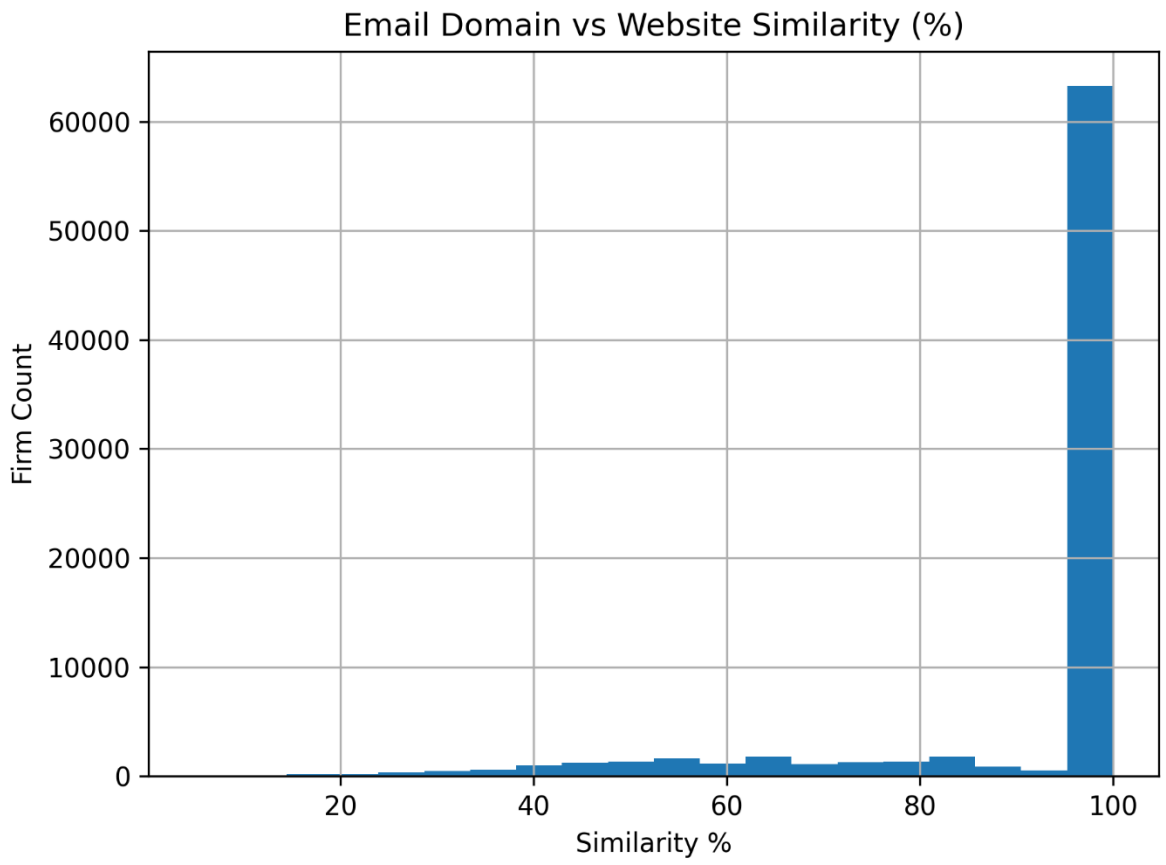

Figure 9: Email Domain Website Root Similarity

A clear trend showed that the majority of firms used a domain that matches their website root entirely with a right skewed distribution.

	FALSE	TRUE
firm_is_multi_domain	46.59624	19.84202
is_shared_infra	31.60136	18.47121
either_flag	61.05751	33.09431
neither_flags	38.94249	66.90569
both_flags	17.14009	5.218919

Table 3: Corporate Structure to Missing Domain Template

For the cases that do not match, firm corporate structure flags were investigated to see if they influenced the difference between website root and email domain. A clear trend was identified between multi domain and shared infrastructure firms and website to domain mapping ambiguity. This influences our approach to predicting email domains for unseen firms towards a RAG and LLM approach, where unseen firms that cannot be retrieved from the fuzzy lookup will be

passed to an LLM to retrieve a website root or email domain from external sources.

6. Imputation Strategy for Missing Emails

To try and recover some of the missing dataset, template consistency was investigated to find firms with 100% consistency as candidates for imputation. However from these consistent firms, steps were taken to ensure unfair assumptions weren't made regarding the template diversity. For example, if a firm has only two entries—one with a missing email—100% consistency isn't meaningful. We cannot safely assume that that same consistency would hold true if the email were recovered.

Firm	Missing Email Count	Non Missing Email Count	Total Count	Missing Email %
merrill lynch, pierce, fenner & smith	726	3	729	99.58848
fidelity personal and workplace advisors	518	7	525	98.66667
beacon pointe wealth advisors	297	37	334	88.92216
morgan stanley wealth management	177	18	195	90.76923
avantax advisory services	155	9	164	94.5122
bernstein private wealth management	154	65	219	70.31963
sequoia financial group	145	6	151	96.02649
cash balance retirement plan of brown brothers harriman & co	136	4	140	97.14286
thrivent investment management	123	6	129	95.34884
private advisor group	120	8	128	93.75
kestra advisory services	91	4	95	95.78947
kovitz investment group partners	91	60	151	60.2649
surevest private wealth	89	2	91	97.8022

apollo asset management	76	10	86	88.37209
koda capital	62	25	87	71.26437
global retirement partners	59	8	67	88.0597
jordan family office	59	6	65	90.76923
industrial and commercial bank of china	58	2	60	96.66667
ashton thomas private wealth	55	9	64	85.9375
independent financial group	54	3	57	94.73684

Table 4: Email Missingness per Firm

To combat this missingness of emails was investigated within the original raw dataset.

	missing_email_count	non_missing_email_count	total_count	missing_email_pct
count	10688	10688	10688	10688
mean	2.203593	3.29753	5.501123	26.29459
std	10.88211	4.607631	12.18476	31.115
min	0	1	1	0
25%	0	1	2	0
50%	0	2	3	0
75%	2	4	6	50
max	726	111	729	99.58848

Table 5: Email Missingness per Firm Statistics

Statistics were gathered, highlighting plenty of cases where imputation would not be suitable. For that reason, a threshold of at least 5 investors and at most 40% missingness were chosen to ensure that we are not imputing emails for firms that we do not know enough about.

7. Candidate Templates and Feature Matrix Design

Following pattern mining and EDA, a structed, enriched candidate template set was assembled to serve as the backbone for template prediction. Using the full set of mined email template sequences, a master list of 404 unique structural

templates were gathered and saved to the database. Each template was decoded into a human-readable format and assigned a unique ID.

To enhance their predictive power, each template was enriched with structural and rule mining features. Structural features include number of tokens, use of initials or middle names or other name structures that are present in the sequence. Mining rule features are derived from the TRuleGrowth results that match the candidate sequence, allowing us to attach a support and confidence values for matching sequences.

Together this will be used to form our feature matrix, where each row is an (‘investor’, ‘firm’ or ‘domain’, ‘candidate template’) triple. Each row is further enriched with features from the investor name that align with the candidate template along with any other coverage that those idiosyncrasies have with those templates – a similar process is applied to firm-level coverage, using an SQL table that maps firms to their respective used templates and diversity.

This design supports pointwise ranking, where LightGBM scores each candidate independently and selects the most probable template based on engineered features.