FGV DIREITO RIO

Diferenciando normas abstratas de atos administrativos concretos com auxílio de ferramentas de machine learning

Pesquisa Empírica em Direito da Regulação e Política Regulatória (EPED 2025)

Lucas Thevenard Gomes **(FGV Rio Law / Regulação em Números)**

Com Camila de Oliveira Lopes; José Luiz Nunes; Simone Diniz Junqueira Barbosa

Agenda

- Por que utilizar **ML** para estudar a **linguagem regulatória**?
- Por que **simplificação regulatória** importa?
- O gargalo informacional no **DOU**
- **Dados, métodos e experimentos**
- Resultados, limitações e **implicações de design**
- Próximos passos

Por que utilizar ML para estudar a linguagem regulatória?

- Técnicas de NLP e ML permitem analisar grandes volumes de texto
  - Crescimento e importância das normas secundárias no ordenamento jurídico
- Aplicações regulatórias:
  - Extração de entidades e relações (exemplo: pesquisa sobre o papel da Anvisa)
  - Análise do processo regulatório (exemplo: monitoramento de consultas públicas e AIRs)
  - **Análise da complexidade linguística da regulação**

Simplificação regulatória: por quê?

- **+ complexidade** = **+ custos de conformidade** e barreiras de acesso
- **Outros benefícios**: transparência**, previsibilidade e enforcement mais eficiente (segurança jurídica)
- Linguagem simples como diretriz de **boa governança**
  - Estratégia Regula Melhor
  - PL 4.401/2024

Complexidade normativa como uma linha de pesquisa inexplorada no Brasil

- Quão complexa é a linguagem regulatória?
- Quais fatores influenciam essa complexidade e por que ela persiste?
- Quais os efeitos (sociais e institucionais) da complexidade normativa?

Por que a complexidade persiste?

- Iniciativas de “plain language” nos EUA têm efeitos limitados (estudo de corpus 1951–2009)
  - Magic Spell ou Copia e Cola?
  - Evidências apontam para uma complexidade intencional (efeito “performativo” ou "magic spell")
- Incentivos institucionais podem favorecer textos densos
- Hipótese estratégica: complexidade como mecanismo de controle do processo

O gargalo no Brasil (DOU, Seção 1)

- Ausência de uma tipologia adequada de atos administrativos
  - Mistura de **atos normativos-abstratos** e **atos concretos**
- Ausência de marcador oficial → alto custo de triagem
- Consequência: “achar a agulha no palheiro”
- Obs: Decreto de consolidação dos atos normativos inferiores a decreto não resolveu o problema

Exemplo motivador: transporte de materiais inflamáveis

- Busca temática retorna tanto regras gerais quanto autorizações individuais
- Usuário precisa descobrir **quem regula** e **o que é norma**
- Meta: reduzir ruído e ordenar por relevância normativa

Objetivo da pesquisa

- Classificar automaticamente **normas abstratas** vs **atos concretos** publicados por agências federais
- Compreender os aspectos linguísticos que diferenciam esses atos
- Pré-requisito para usar dados textuais do DOU em larga escala

Dados

- DOU – Seção 1; 10 agências federais independentes (2018–2022)
- Base inicial: **58.694** atos

Pré-processamento crítico

- Presença de **artigos** como indício forte de normatividade (suficiente, não necessário)
- **50.139** sem artigos; amostra (n=100) → todos concretos
- Remoções: documentos fora de escopo / casos ambíguos
- Base final para modelagem: **8.555** atos

Representação e modelos

- **TF-IDF** (n-grams 1–2, ~5k features)
- **Embeddings BERT** (BERTimbau) e **LEGAL-BERT** (pt-br)
- Algoritmos: KNN, DT, RF, MLP, AdaBoost, NB, SVM, LogReg (com *grid search*)

Experimento 1 – Intra-agência (previsão temporal)

- Treina em anos passados; prediz ano futuro da **mesma agência**
- Acurácia típica: **90–96%** com modelos clássicos, em ambas as classes
- Aprendizado “dentro de casa” é relativamente fácil

Experimento 2 – Generalização (TF-IDF)

- Treino: várias agências; teste **fora da amostra** (ex.: ANVISA/ANEEL)
- Melhor DT: ~**82%** (abstratos) | ~**75%** (concretos)
- Dependência de poucas *features* (≈12 termos)

Experimento 2 – Generalização TF-IDF (out-of-sample)

Model	Accuracy (%)		F1-Score
	Abstract	Concrete	
Nearest Neighbors	95.06%	55.93%	0.55
Decision Tree	81.95%	75.09%	0.61
Random Forest	87.66%	47.35%	0.48
Neural Net	99.35%	10.76%	0.39
AdaBoost	92.21%	14.73%	0.38
Naive Bayes	97.40%	3.59%	0.37
SVM Linear	98.83%	41.17%	0.50
SVM RBF	99.74%	16.75%	0.41
SVM Sigmoid	98.83%	50.74%	0.54
Logistic Regression	99.22%	17.74%	0.41

Experimento 3 – Embeddings BERT

- Desempenho **comparável** ao TF-IDF
- Embeddings gerais não capturam nuances jurídico-regulatórias específicas
- Trade-off: robustez semântica vs. *signal* estilístico

Experimento 3 – Embeddings BERT (out-of-sample)

Model	Accuracy (%)		F1-Score
	Abstract	Concrete	
Nearest Neighbors	95.71%	41.51%	0.48
Decision Tree	81.69%	19.34%	0.36
Random Forest	97.79%	4.58%	0.37
Neural Network	97.14%	23.46%	0.42
AdaBoost	91.43%	45.10%	0.48
Naive Bayes	95.58%	27.93%	0.43
SVM (Linear)	94.55%	45.33%	0.50
SVM (RBF)	97.40%	12.55%	0.39
SVM (Sigmoid)	97.53%	11.90%	0.39
Logistic Regression	94.55%	37.28%	0.46

Experimento 4 – LEGAL-BERT (fine-tuning)

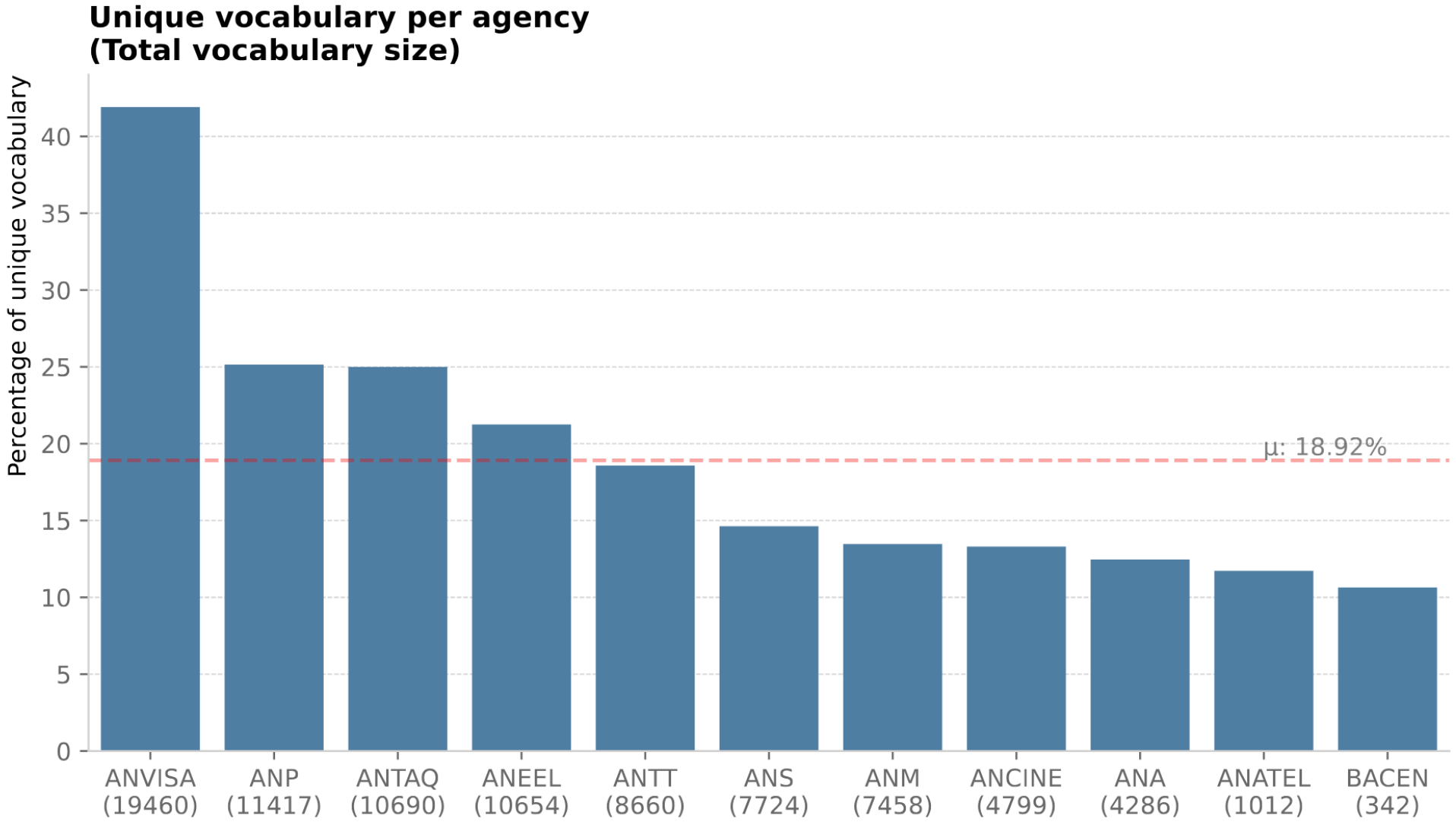
- **Abstratos:** ~**92,85%**
- **Concretos:** ~**51,74%** (desafio persistente)
- Ganho em normas gerais; queda em concretos fora do domínio

Experimento 4 – LEGAL-BERT com fine-tuning (out-of-sample)

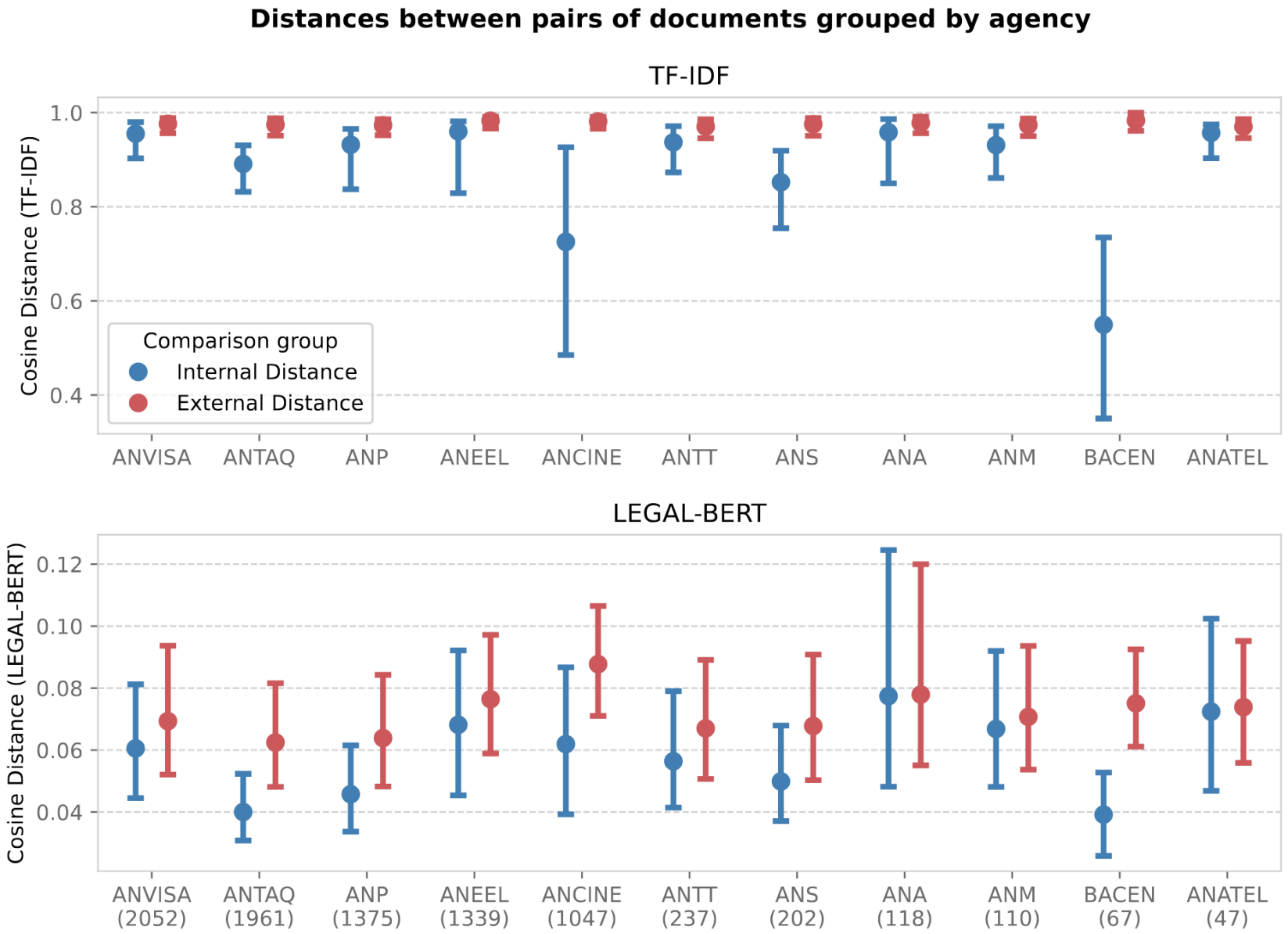
Model	Accuracy (%)		F1-Score
	Abstract	Concrete	
Nearest Neighbors	95.19%	35.90%	0.46
Decision Tree	85.06%	24.34%	0.38
Random Forest	98.70%	14.92%	0.40
Neural Network	97.66%	28.73%	0.44
AdaBoost	90.00%	38.00%	0.45
Naive Bayes	88.18%	37.28%	0.48
SVM (Linear)	96.36%	38.65%	0.43
SVM (RBF)	98.31%	23.96%	0.43
SVM (Sigmoid)	87.40%	30.22%	0.41
Logistic Regression	95.06%	40.48%	0.48

Por que generalizar é difícil?

- Diferenças **linguísticas e estilísticas** entre agências
- Concretos são mais **variáveis** e menos padronizados
- Hipóteses: marcadores de agência/tema vs. diversidade de formatos

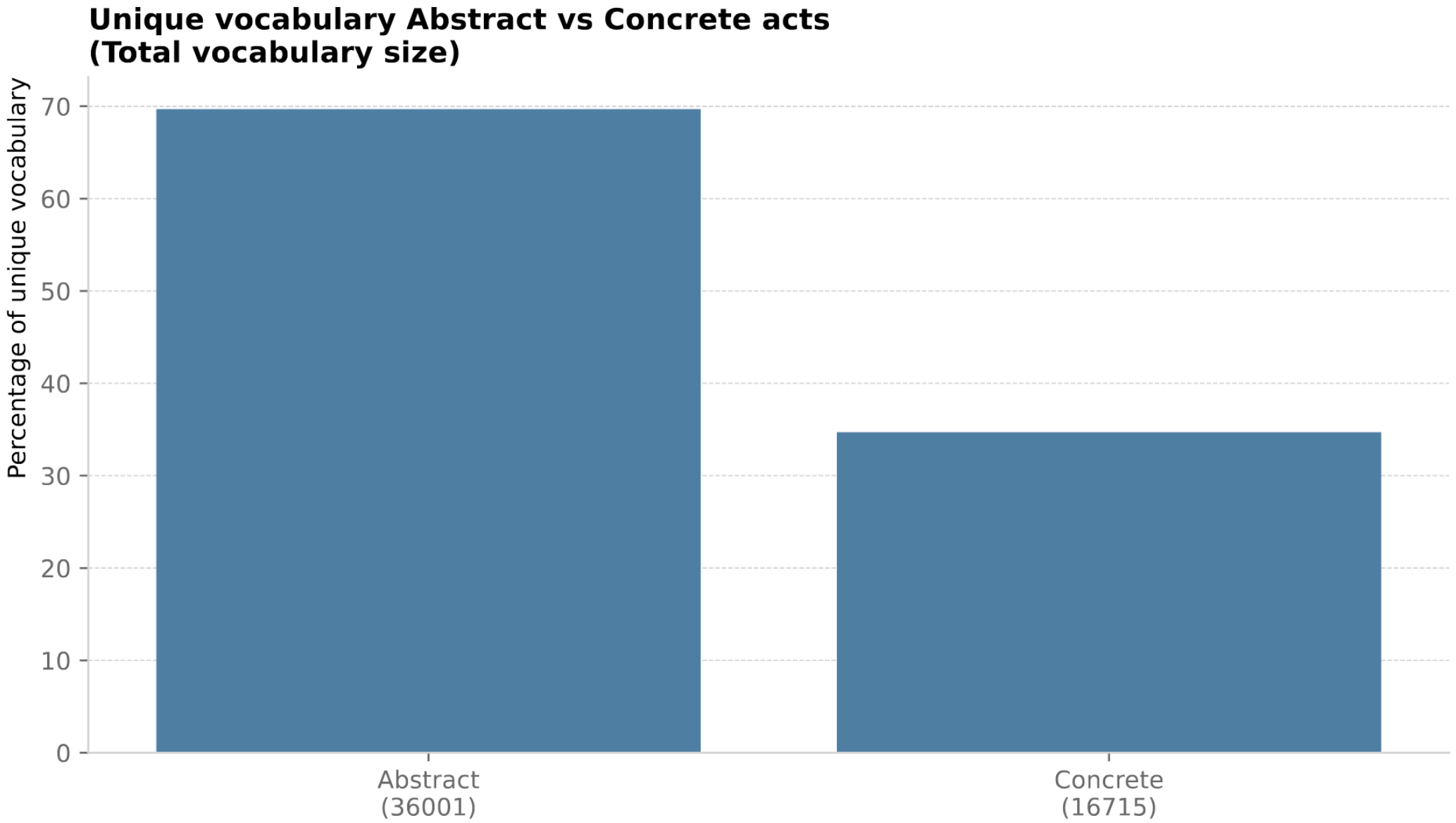

Unique vocabulary per agency
(Total vocabulary size)
Percentage of unique vocabulary
0
5
10
15
20
25
30
35
40
μ: 18.92%
ANVISA (19460)
ANP (11417)
ANTAQ (10690)
ANEEL (10654)
ANTT (8660)
ANS (7724)
ANM (7458)
ANCINE (4799)
ANA (4286)
ANATEL (1012)
BACEN (342)

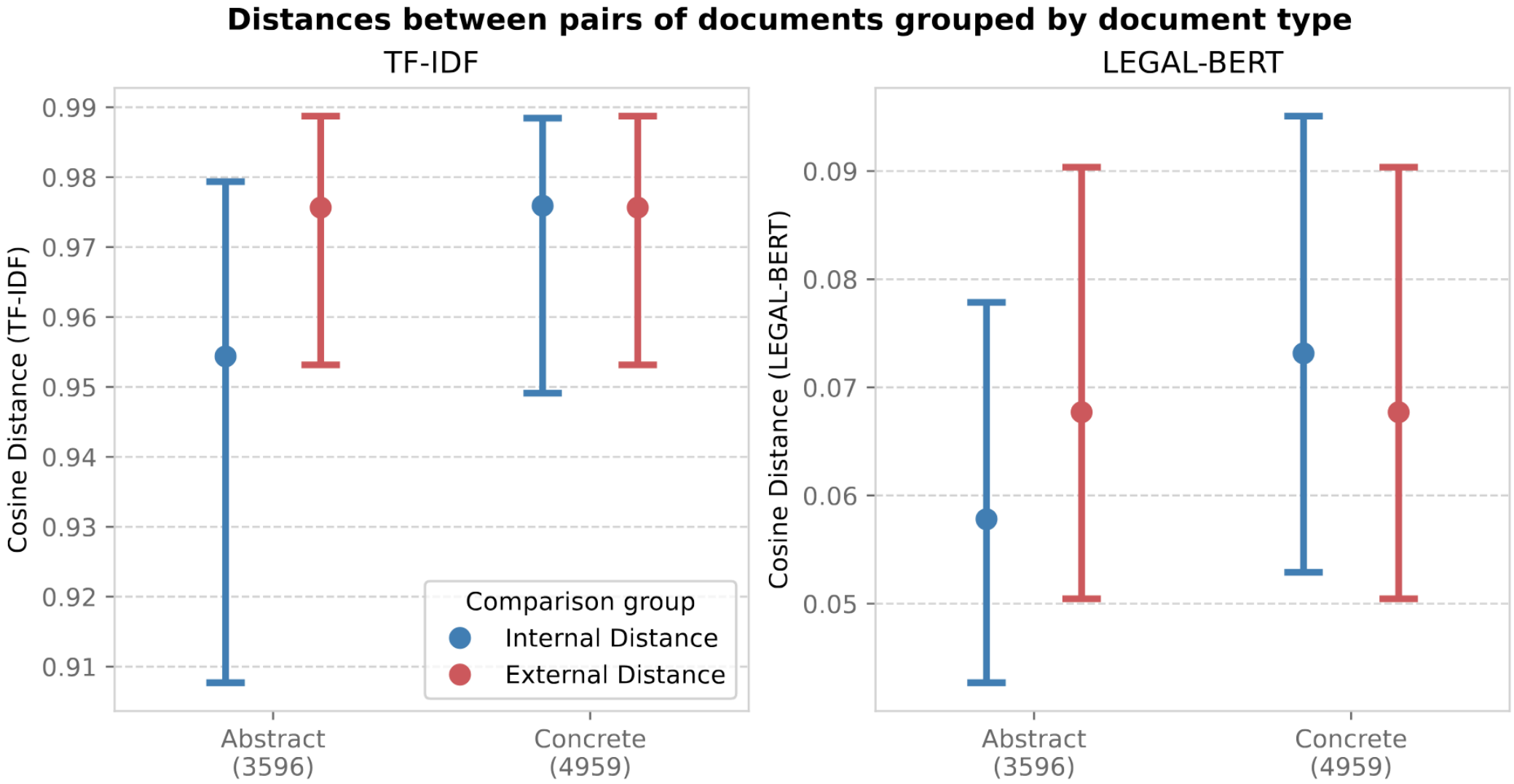
Diferenciando normas abstratas de atos administrativos concretos



Diferenciando normas abstratas de atos administrativos concretos

FGV DIREITO RIO




Distances between pairs of documents grouped by document type
TF-IDF
LEGAL-BERT
Cosine Distance (TF-IDF)
Cosine Distance (LEGAL-BERT)
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.09
0.08
0.07
0.06
0.05
Comparison group
Internal Distance
External Distance
Abstract
(3596)
Concrete
(4959)

Limitação técnica relevante e próximos passos

- Limite de **512 tokens** em (LEGAL-)BERT atinge parte do corpus
- Alternativas: modelos de **contexto longo** e **segmentação por artigos**
- Próximo passo: *hierarchical transformers* e sumarização estruturada

Limitações & próximos passos

- **Generalização** entre agências e domínios
- **Contexto longo** e segmentação por artigos
- Avaliação com **usuários** (tarefa, tempo, *precision/recall* percebido)

Conclusões

- **(i)** Pré-classificação já **reduz custo cognitivo** e habilita melhores portais de informação
- **(ii)** Classificação **intra-agência** é viável e relativamente simples, podendo gerar ganhos de informação relevantes (exemplo da ANEEL)
- **(iii)** Generalização inter-agências é o grande desafio atual
- **(iv)** A fronteira desta pesquisa está na promoção de uma melhor compreensão dos elementos linguísticos (vocabulário, estruturas típicas, tarefas regulatórias usuais) que diferenciam atos normativos e concretos

Obrigado!

Contato: lucas.thevenard@fgv.br

Projeto: Regulação em Números