Who gets heard?

Language, interests, and the influence of textual public input in the regulatory decision-making process of the Brazillian telecommunications sector

Lucas Thevenard

Motivation

- Anvisa and Aneel's public consultations
  - Do participants' need to "speak the language" of the agency to get heard?
- A new way to analyse the content of contributions
  - Text mining and machine learning models of textual analysis

FGV DIREITO RIO

Participation assymetry in the literature

- Evidence that some groups (most notably, the business sector) are overrepresented in public consultations
  - Coglianese, 2006; Yackee, 2006; Silva, 2012, Salinas *et al.*, 2020.
- Interest group theories
  - Capture theory and the Iron Triangle
  - "Hollow core", ACF and policy networks

Why Anatel?

- High level of transparency and accessible data.
- Full text of contributions, responses, and participant IDs.
- Strategic sector: digital communications and telecom.

Three hypothesis of this study

Hypothesis 1: Contributions with similar textual content receive similar regulatory responses.

Hypothesis 2: Contributions that are textually closer to the agency’s discourse are more likely to be accepted.

Hypothesis 3: Once textual content is accounted for, group affiliation loses explanatory power.

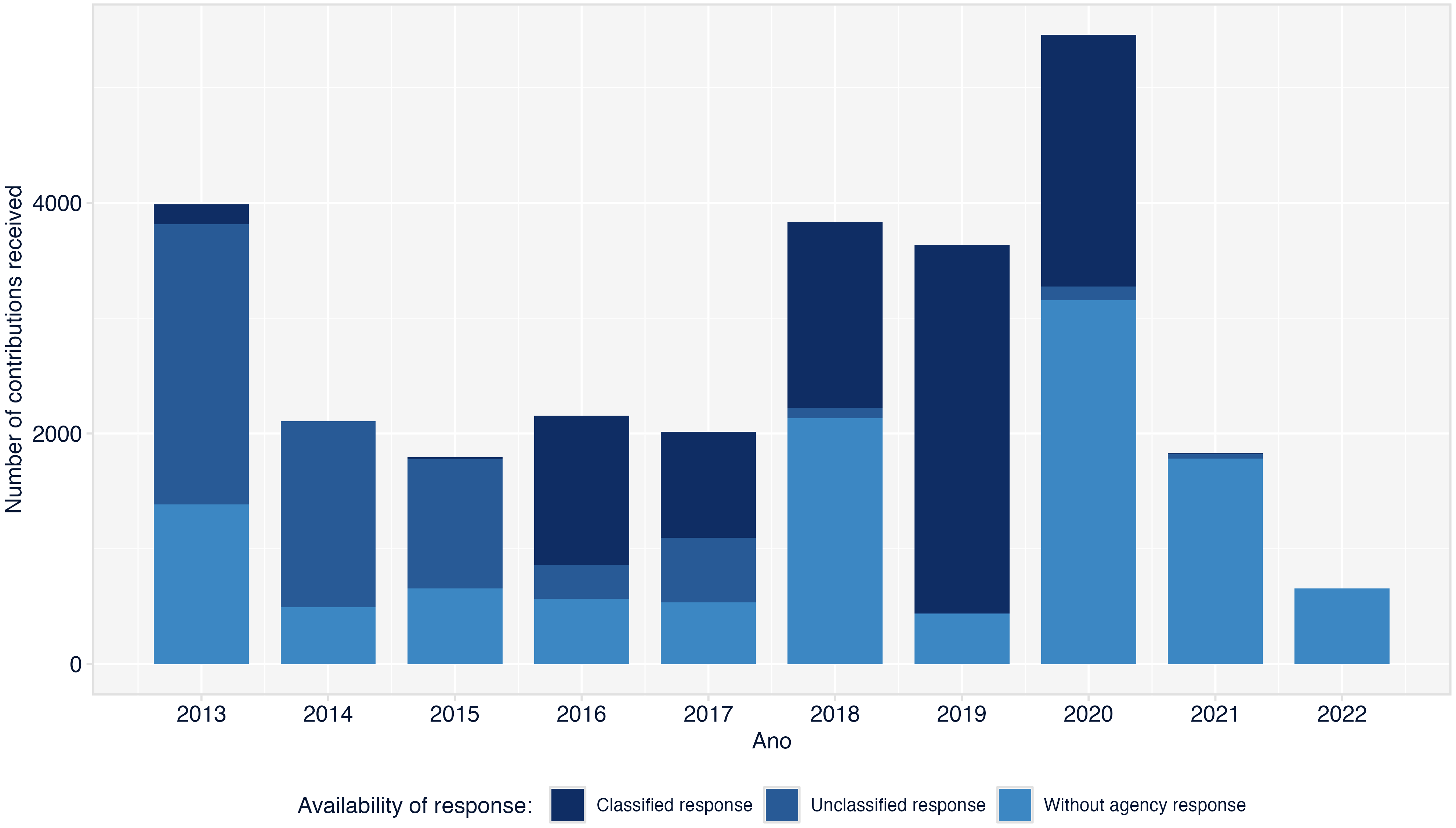
FGV DIREITO RIO

Data collection

- 1,456 consultations from 1999 to 2022
  - Only 488 consultations received formal, itemized responses from the agency.
- From 2013 onward, of the 15,679 contributions that received a response, 9,404 (approximately 60%) were accompanied by a classification.

Who gets heard?

FGV DIREITO RIO



SOURCE: Created by the author, based on data from Anatel

Final universe

- 9,404 contributions received by Anatel since 2013 that were accompanied by a classified response.
- Simplification of impact categories:
    - Accepted, partially accepted, not accepted, and not applicable ---> 0 x 1
- Classification of Interest group affiliation:
    - (1) regulated economic agents, (2) non-regulated economic agents, (3) sector professionals, (4) amateur radio operators, (5) interested individuals, and (6) others.

Limitations of the data

- Reliability
- Representativeness
- Validity

Text Representation

FGV DIREITO RIO

What is TF-IDF?

TF-IDF (Term Frequency–Inverse Document Frequency) is a numerical statistic used in text mining and Natural Language Processing (NLP) to reflect how important a word is to a document in a collection or corpus.

It combines two metrics: Term Frequency (TF) and Inverse Document Frequency (IDF).

1. Term Frequency (TF)

Measures how frequently a term appears in a document.

$$\mathrm{TF}(t, d) = \frac{f_{t,d}}{\sum_k f_{k,d}}$$

Where:

- $(f_{t,d})$ is the number of times term (t) appears in document (d)
- The denominator is the total number of terms in document (d)

FGV DIREITO RIO

2. Inverse Document Frequency (IDF)

Measures how important a term is by reducing the weight of terms that appear in many documents.

$$\mathrm{IDF}(t, D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)$$

Where:

- (N) is the total number of documents in the corpus (D)
- $(|\{d \in D : t \in d\}|)$ is the number of documents where term (t) appears

TF-IDF Formula

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

This score increases with the frequency of a term in a document but is offset by how common the term is across the corpus.

- **Intuition**:
  - Common words (like “the”, “and”) get low scores.
  - Terms that are frequent in a document but rare across the corpus get high scores.
  - Documents (contributions) become vectors in a high-dimensional space, where each dimension corresponds to a term in the vocabulary.

Predictive modelling

- Open code is available for all the modelling stages.

Table 4 – Consolidated predictive performance (test and validation)

Model	Evaluation	Accuracy	ROC-AUC	Precision	Recall	F1	N
KNN	Test	85.25%	0.91	0.83	0.80	0.81	1,505
	Validation	84.90%	0.91	0.83	0.79	0.81	1,881
RF	Test	86.05%	0.94	0.87	0.78	0.81	1,505
	Validation	86.76%	0.94	0.87	0.80	0.82	1,881
SVC	Test	88.24%	0.94	0.89	0.82	0.84	1,505
	Validation	88.57%	0.94	0.89	0.83	0.85	1,881

Explanatory modeling (Model 1)

- Cosine distance as a textual distance metric

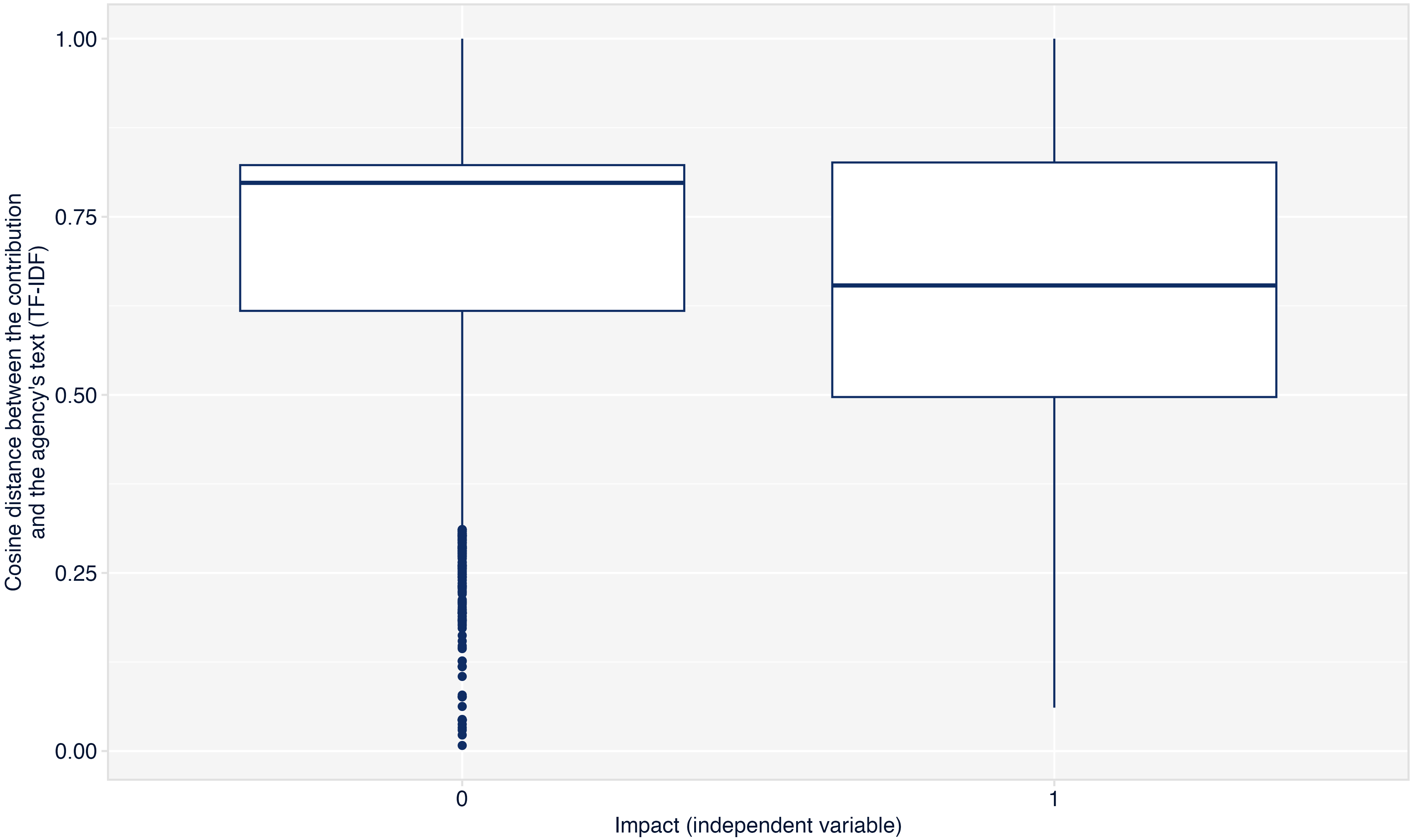
FGV DIREITO RIO

Explanatory modeling (Model 1)

- Cosine distance as a textual distance metric

Table 6 – Logistic regression: impact explained by textual distance and year of consultation

	Coefficient	Std. Error	z-value	p-value	
Intercept	182,02	30,02	6,05	1,44E-09	***
Cosine distance	-1,77	0,12	-15,13	2,00E-16	***
Year	-0,09	0,01	-6,04	1,56E-09	***
Significance codes: 0 *** 0,001 ** 0,01 * 0,05 . 0,1					



1.00
0.75
0.50
0.25
0.00
Cosine distance between the contribution and the agency's text (TF-IDF)
0
1
Impact (independent variable)

FGV DIREITO RIO

Explanatory modeling (Model 2)

Table 7 – Logistic regression: impact explained by model-based probabilities and consultation year

	Coefficient	Std. Error	z-value	p-value	
Intercept	54,71	68,99	0,79	0,4277	
Year	-0,03	0,03	-0,87	0,3820	
KNN Probability	-0,71	0,31	-2,32	0,02	*
RF Probability	11,26	0,48	23,51	2,00E-16	***
SVC Probability	2,15	0,33	6,36	2,04E-10	***
	Significance codes: 0 *** 0,001 ** 0,01 * 0,05 . 0,1				

FGV DIREITO RIO

Table 8 –Logistic Regression: Impact explained by year, ML scores, and interest group category

	Coefficient	Std. Error	z-value	p-value	
Intercept	9,70	69,39	0,14	0,8887	
Year	-0,01	0,03	-0,22	0,8282	
KNN Probability	-0,68	0,31	-2,21	0,0269	*
RF Probability	11,62	0,52	22,55	2,00E-16	***
SVC Probability	1,94	0,35	5,55	2,79E-08	***
Participant: Regulated Agent	-0,37	0,13	-2,93	0,0034	**
Participant: Non-Regulated Agent	-0,55	0,26	-2,08	0,0376	*
Participant: Sector Professional	-3,04	0,50	-6,10	1,05E-09	***
Participant: Amateur Radio Operator	-1,67	0,60	-2,75	0,0059	**
Participant: Other	-0,69	0,3	-2,33	0,0197	*

Significance codes: 0 *** 0,001 ** 0,01 * 0,05 . 0,1

FGV DIREITO RIO

Going back to our hypothesis

Hypothesis 1: Contributions with similar textual content receive similar regulatory responses.

Hypothesis 2: Contributions that are textually closer to the agency's discourse are more likely to be accepted.

Hypothesis 3: Once textual content is accounted for, group affiliation loses explanatory power.

FGV DIREITO RIO

Some additional analysis

- Is it technical language that matters?
- Repetition and impact: which groups are more likely to repeat their arguments?
- Group cohesion: internal and external textual distances across interest groups

FGV DIREITO RIO

Thank You.