Data: 22/01/2024 – Eduardo Corrêa

1. Estatísticas dos Labels da Base Concatenada: train + validation + dev *versus* Base train apenas

- Após liberarem os labels da base "dev", eu concatenei as 3 bases de dados em um único arquivo e gerei as estatísticas abaixo, onde:
  - N = total de instâncias
  - q = total de labels
  - LCard = label cardinality, que é a média de labels por instância
  - LDens = label density, que é LCard / N
  - NDist = total de combinações de labels distintas que ocorrem no BD
  - NUniq = total de combinações únicas, ou seja, que ocorrem em apenas 1 instância
  - PUniq = proporção de combinações únicas
    - 55% das instâncias tem uma combinação "só dela".
  - NMax = frequência do labelset que mais ocorre na base de dados
  - max_labelsets = quem são os labelsets com NMax
    - Na verdade o cara mais frequente é o "sem labels" (não é uma propaganda), em 1508 instâncias, o que dá 17,7%
  - label_freqs = frequência de cada label
  - label_rank = labels ordenados por frequência (do que mais ocorre no BD para o que menos ocorre).

- Na primeira coluna, coloquei as estatísticas apenas da base de treino e na segunda das 3 bases concatenadas.
  - A estatística que mais mudou foi NDist
    - A quantidade de combinações distintas de labels, que subiu de 812 para 936 – o que é uma boa justificativa para usar o BR.

- Na última página ordenei os labels por frequência (do mais frequente para o menos frequente). A ordem é quase igual na base de treino e no BD concatenado, há apenas uma troca de posição entre os labels 'Exaggeration/Minimisation' e 'Appeal to fear/prejudice'.

	Train	Train + Valid + Dev
N	7000	8500 (7000 train + 500 valid + 1000 dev)
q	20	20
LCard	1,63	1,64
LDens	0,08064286	0,081988235
NDist	812	936
NUniq	452	513
PUniq	0,55665025	0,548076923
NMax	1264	1508
max_labelsets	[()]	[()]
label_freqs	'Appeal to authority': 850, 'Appeal to fear/prejudice': 337, 'Bandwagon': 97, 'Black-and-white Fallacy/Dictatorship': 780, 'Causal Oversimplification': 240, 'Doubt': 350, 'Exaggeration/Minimisation': 356, 'Flag-waving': 571, 'Glittering generalities (Virtue)': 488, 'Loaded Language': 1750, "Misrepresentation of Someone's Position (Straw Man)": 62, 'Name calling/Labeling': 1518, 'Obfuscation, Intentional vagueness, Confusion': **21**, 'Presenting Irrelevant Data (Red Herring)': 59, 'Reductio ad hitlerum': 63, 'Repetition': 305, 'Slogans': 667, 'Smears': **1990**, 'Thought-terminating cliché': 528, 'Whataboutism': 258	'Appeal to authority': 1049, 'Appeal to fear/prejudice': 430, 'Bandwagon': 120, 'Black-and-white Fallacy/Dictatorship': 931, 'Causal Oversimplification': 314, 'Doubt': 419, 'Exaggeration/Minimisation': 445, 'Flag-waving': 702, 'Glittering generalities (Virtue)': 595, 'Loaded Language': 2188, "Misrepresentation of Someone's Position (Straw Man)": 76, 'Name calling/Labeling': 1896, 'Obfuscation, Intentional vagueness, Confusion': 31, 'Presenting Irrelevant Data (Red Herring)': 73, 'Reductio ad hitlerum': 78, 'Repetition': 374, 'Slogans': 828, 'Smears': 2414, 'Thought-terminating cliché': 644, 'Whataboutism': 331

label_rank (ordenado pela freq. descendente)	'Smears': **1990**, 'Loaded Language': 1750, 'Name calling/Labeling': 1518, 'Appeal to authority': 850, 'Black-and-white Fallacy/Dictatorship': 780, 'Slogans': 667, 'Flag-waving': 571, 'Thought-terminating cliché': 528, 'Glittering generalities (Virtue)': 488, 'Appeal to fear/prejudice': 337, 'Exaggeration/Minimisation': 356, 'Doubt': 350, 'Repetition': 305, 'Whataboutism': 258, 'Causal Oversimplification': 240, 'Bandwagon': 97, 'Reductio ad hitlerum': 63, "Misrepresentation of Someone's Position (Straw Man)": 62, 'Presenting Irrelevant Data (Red Herring)': 59, 'Obfuscation, Intentional vagueness, Confusion': **21**,	'Smears': 2414, 'Loaded Language': 2188, 'Name calling/Labeling': 1896, 'Appeal to authority': 1049, 'Black-and-white Fallacy/Dictatorship': 931, 'Slogans': 828, 'Flag-waving': 702, 'Thought-terminating cliché': 644, 'Glittering generalities (Virtue)': 595, 'Exaggeration/Minimisation': 445, 'Appeal to fear/prejudice': 430, 'Doubt': 419, 'Repetition': 374, 'Whataboutism': 331 'Causal Oversimplification': 314, 'Bandwagon': 120, 'Reductio ad hitlerum': 78, "Misrepresentation of Someone's Position (Straw Man)": 76, 'Presenting Irrelevant Data (Red Herring)': 73, 'Obfuscation, Intentional vagueness, Confusion': 31,

2. Estatísticas dos Labels: Base validation *versus* Base dev

- Na página a seguir, as estatísticas das bases validation (1ª coluna) e dev (2ª coluna)
  - O PUniq dá uma amentada
  - A proporção de instâncias sem rótulo (ou seja, que não são propagandas) é mais ou menos a mesma do BD de treino.

	Validation	Dev
N	500	1000
q	20	20
LCard	1,68	1,81
LDens	0,0839	0,09045
NDist	173	301
NUniq	120	189
PUniq	0,693641618	0,627906977
NMax	88	156
max_labelsets	[()]	[()]
label_freqs	'Appeal to authority': 63, 'Appeal to fear/prejudice': 27, 'Bandwagon': 7, 'Black-and-white Fallacy/Dictatorship': 53, 'Causal Oversimplification': 21, 'Doubt': 24, 'Exaggeration/Minimisation': 27, 'Flag-waving': 42, 'Glittering generalities (Virtue)': 36, 'Loaded Language': 135, "Misrepresentation of Someone's Position (Straw Man)": 4, 'Name calling/Labeling': 116, 'Obfuscation, Intentional vagueness, Confusion': 2, 'Presenting Irrelevant Data (Red Herring)': 4, 'Reductio ad hitlerum': 4, 'Repetition': 23, 'Slogans': 50, 'Smears': 142, 'Thought-terminating cliché': 38, 'Whataboutism': 21	'Appeal to authority': 136, 'Appeal to fear/prejudice': 66, 'Bandwagon': 16, 'Black-and-white Fallacy/Dictatorship': 98, 'Causal Oversimplification': 53, 'Doubt': 45, 'Exaggeration/Minimisation': 62, 'Flag-waving': 89, 'Glittering generalities (Virtue)': 71, 'Loaded Language': 303, "Misrepresentation of Someone's Position (Straw Man)": 10, 'Name calling/Labeling': 262, 'Obfuscation, Intentional vagueness, Confusion': 8, 'Presenting Irrelevant Data (Red Herring)': 10, 'Reductio ad hitlerum': 11, 'Repetition': 46, 'Slogans': 111, 'Smears': 282, 'Thought-terminating cliché': 78, 'Whataboutism': 52

label_rank (ordenado pela freq. descendente)	'Smears': 142, 'Loaded Language': 135, 'Name calling/Labeling': 116, 'Appeal to authority': 63, 'Black-and-white Fallacy/Dictatorship': 53, 'Slogans': 50, 'Flag-waving': 42, 'Thought-terminating cliché': 38, 'Glittering generalities (Virtue)': 36, 'Appeal to fear/prejudice': 27, 'Exaggeration/Minimisation': 27, 'Doubt': 24, 'Repetition': 23, 'Whataboutism': 21 'Causal Oversimplification': 21, 'Bandwagon': 7, "Misrepresentation of Someone's Position (Straw Man)": 4, 'Presenting Irrelevant Data (Red Herring)': 4, 'Reductio ad hitlerum': 4, 'Obfuscation, Intentional vagueness, Confusion': 2,	'Loaded Language': 303, 'Smears': 282, 'Name calling/Labeling': 262, 'Appeal to authority': 136, 'Slogans': 111, 'Black-and-white Fallacy/Dictatorship': 98, 'Flag-waving': 89, 'Thought-terminating cliché': 78, 'Glittering generalities (Virtue)': 71, 'Appeal to fear/prejudice': 66, 'Exaggeration/Minimisation': 62, 'Causal Oversimplification': 53, 'Whataboutism': 52 'Repetition': 46, 'Doubt': 45, 'Bandwagon': 16, 'Reductio ad hitlerum': 11, "Misrepresentation of Someone's Position (Straw Man)": 10, 'Presenting Irrelevant Data (Red Herring)': 10, 'Obfuscation, Intentional vagueness, Confusion': 8,