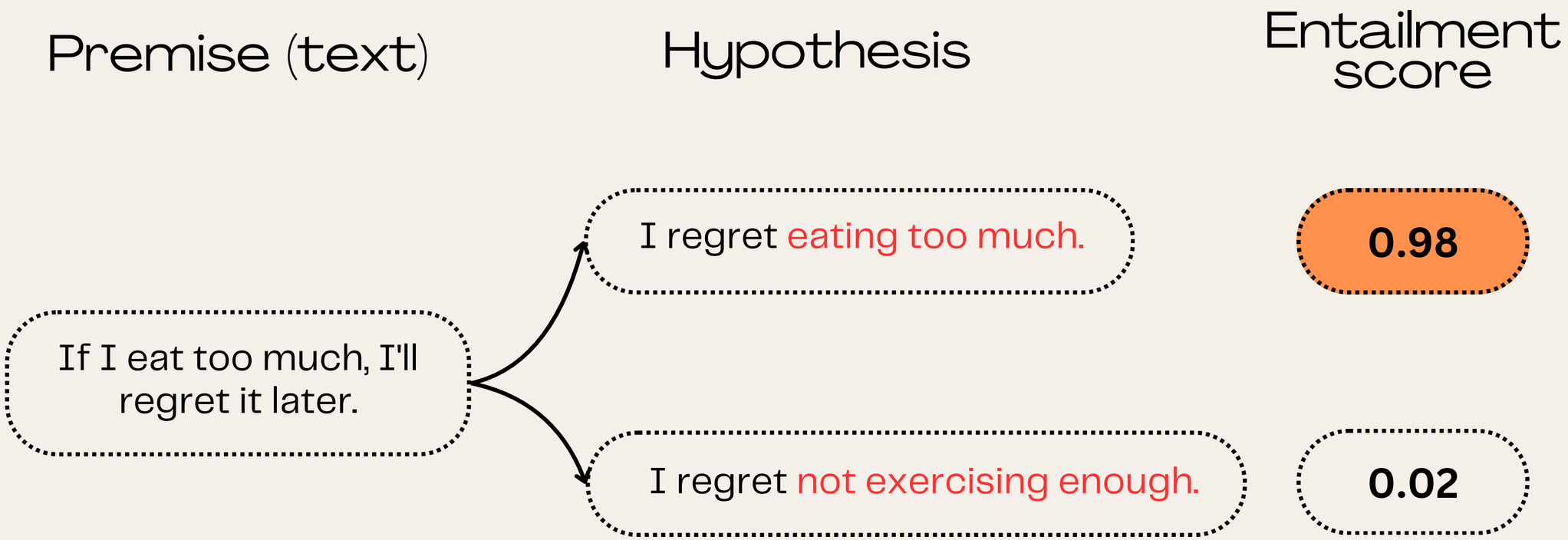
Entailment Results

rayane.ghilene@ensea.fr

Entailment-based zero-shot classification.



Experimental results:

Model: roberta-large-mnli
Template: "This text contains {} speech"
Labels: ***neutral*** instead of neither and ***hateful*** instead of hate

roBERTa													
	all-10%	hateval	Davidson	Founta	fox	gab	grimminger	hasoc2019	hasoc 2020	olid	reddit	stormfront	trac	AVG
BERT F1 Score (macro)		63,6	73	70,1	47,8	87,5	51,9	32,9	41,7	65,6	81,7	66,9	67,1	62,4833
NEW F1 Score (macro)		**61,4**	**44,7**	**57,5**	**55,2**	**67,1**	**56,1**	**30,9**	**42,7**	**61,5**	**58**	**62,6**	**64,2**	55,1583
OLD F1 Score (macro)		52,1	33,8	37,5	55,2	55,1	38,7	17,9	26,8	53,2	47,9	45	45,8	42,4166
Improvement %		15,146	24,384	34,78	0	17,88	31,0160	42,0711	37,2365	13,49	17,41	28,115	28,66	23,1001

Experimental results:

Model: bart-large-mnli
Template: "This text contains {} speech"
Labels: ***neutral*** instead of neither and ***hateful*** instead of hate

BART														
	all-10%	hateval	Davidson	Founta	fox	gab	grimminger	hasoc2019	hasoc 2020	olid	reddit	stormfront	trac	AVG
NEW F1 Score (macro)		**52**	**29**	**50,9**	**48,4**	**53,7**	**40,2**	**27,4**	**36**	**58,3**	**43,4**	**53**	**53,9**	45,5166
OLD F1 Score (macro)		57,9	39	49,1	57,9	63,6	48,6	20,2	32	55,4	56,2	60,1	46,2	48,85
Improvement %		-11,346	-34,48	3,536	-19,62	-18,4	-20,8955	26,2773	11,1111	4,974	-29,49	-13,396	14,28	-7,3233

The same template that improves the performance of **roberta** reduces the performance of **bart**

Experimental results:

Model: bart-large-mnli
Template: "This text has {} speech"
Labels: **free** instead of neutral

BART 2													
	all-10%	hateval	Davidson	Founta	fox	gab	grimminger	hasoc2019	hasoc 2020	olid	reddit	stormfront	trac	AVG
BERT F1 Score (macro)		63,6	73	70,1	47,8	87,5	51,9	32,9	41,7	65,6	81,7	66,9	67,1	62,4833
NEW F1 Score (macro)		**59,7**	**47,3**	**57,4**	**54,3**	**62,7**	**52**	**32,8**	**42,6**	**62,4**	**54,9**	**58,2**	**51,9**	53,0166
OLD F1 Score (macro)		57,9	39	49,1	57,9	63,6	48,6	20,2	32	55,4	56,2	60,1	46,2	48,85
Improvement %		3,0150	17,547	14,45	-6,629	-1,43	6,53846	38,414	24,8826	11,21	-2,367	-3,2646	10,98	7,85916

The choice of template has a direct effect on the classification performance and differ between models

Analysis of the effect of the hypothesis template on the classification:

Model: roberta-large-mnli

Dataset: Davidson

Labels: ['offensive' 'ok' 'hateful']

Template	F1 Score
"this text contains {} speech."	**45,7**
"this text conveys {} speech."	**40,8**
"this text reflects {} speech."	38,3
"this text demonstrates {} speech."	35
"this text shows {} speech."	35,1
"this text implies {} speech."	33,2
"this text reveals {} speech."	37,8
"this text exhibits {} speech."	38,8
"this text portrays {} speech."	33
"this text discusses {} speech."	34,8
"this text addresses {} speech."	34,2
"this text illustrates {} speech."	35,9
"this text expresses {} speech."	**44,5**
"this text articulates {} speech."	**45,1**
"this text suggests {} speech."	30,1
"this text narrates {} speech."	**43,2**
"this text questions {} speech."	32,6
"this text highlights {} speech."	36,8
"this text investigates {} speech."	33,8
"this text supports {} speech."	22,6

Analysis of the effect of the hypothesis template on the classification:

Model: roberta-large-mnli

Dataset: Davidson

Labels: ['offensive' 'ok' 'hateful']

Template	F1 Score
"this text contains {} perspective."	**46,6**
"this text contains {} tone."	44,4
"this text contains {} argument."	42,3
"this text contains {} intent."	41,5
"this text contains {} message."	44,5
"this text contains {} opinion."	39,6
"this text contains {} behavior."	41,3
"this text contains {} attitude."	44,2
"this text contains {} topic."	45,3
"this text contains {} issue."	**46,1**
"this text contains {} feeling."	41,3
"this text contains {} idea."	41,7
"this text contains {} viewpoint."	**45,9**
"this text contains {} belief."	42,8
"this text contains {} theme."	44,4
"this text contains {} event."	43,8
"this text contains {} concept."	45,8
"this text contains {} concern."	44,8
"this text contains {} subject."	43
"this text contains {} claim."	43,4