

A APPENDIX

A.1 TEF OPERATION EXAMPLE

Given classifier F , input sample $s = \text{"a poignant comedy that offers food for thought."}$, original attribution scores $A(s, F, l)$, find the adversarial sequence of tokens s_{adv} that minimizes $PCC[A(s, F, l), A(s_{adv}, F, l)]$ such that at most $\rho_{max} = 25\%$ of words are changed, $\arg \max_l F(s, l) = \arg \max_l F(s_{adv}, l) = \mathbf{Pos.}$ and s_{adv} fulfills the locality constraints described in Section 3, namely each replacement is a synonym of the original word (Mrkšić et al., 2016), the replacement word needs to have the same Part Of Speech computed by SpaCy (Honnibal et al., 2020) and stop words can not be replaced.

w_i	0	1	2	3	4	5	6	7	8
s	a	poignant	comedy	that	offers	food	for	thought	.
$A(s)$	0.0	-0.08	0.4	0.05	0.16	0.23	0.03	0.22	0.0

The word importance ranking from Section 3 yields *poignant* and *comedy* (in this order) to be the $\lfloor 9 \cdot 0.25 \rfloor = 2$ most important tokens, therefore the candidate replacements for only those are considered. This results in the following two steps of TEF.

1. Step. Replace the most important word *poignant* with it's best candidate, measured by lowest PCC. This candidate is the word *distressing*.

w_i	0	1	2	3	4	5	6	7	8	PCC
s'	a	heartbreaking	comedy	that	offers	food	for	thought	.	
$A(s')$	0.01	-0.21	0.45	0.05	0.18	0.25	0.04	0.22	0.01	0.98
s'	a	distressing	comedy	that	offers	food	for	thought	.	
$A(s')$	-0.13	0.15	1.18	0.12	-1.4	0.75	-0.14	0.37	0.09	0.44
s'	a	alarm	comedy	that	offers	food	for	thought	.	
$A(s')$			Failed POS-Filter							-
s'	a	agonizing	comedy	that	offers	food	for	thought	.	
$A(s')$			Failed Prediction-Filter							-

2. Step. Replace the second-most important word *comedy* with the best valid candidate, in this case the token *comic*.

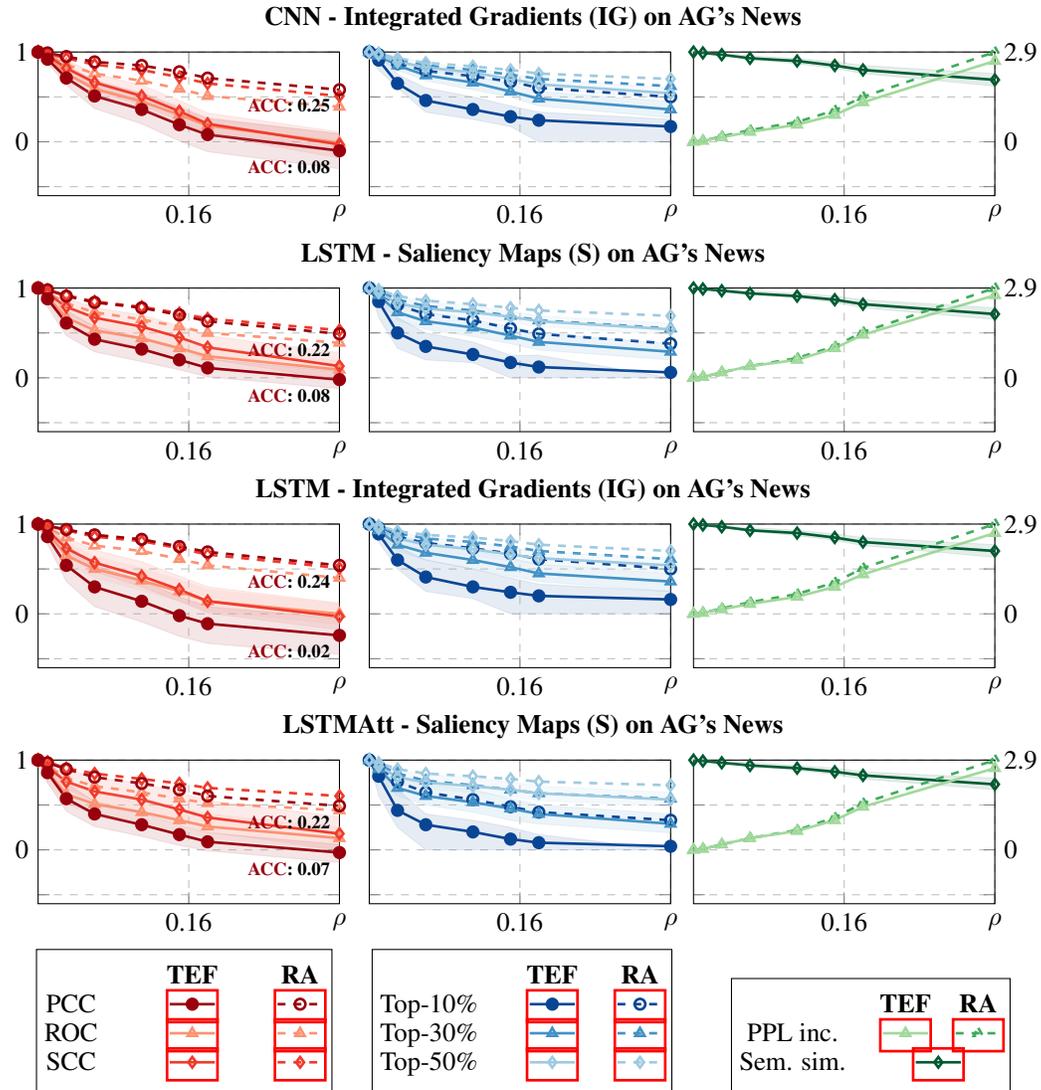
w_i	0	1	2	3	4	5	6	7	8	PCC
s'	a	distressing	humor	that	offers	food	for	thought	.	
$A(s')$	-0.09	0.01	0.34	0.27	0.2	-0.02	0.02	0.27	0.05	0.63
s'	a	distressing	comic	that	offers	food	for	thought	.	
$A(s')$	-0.02	0.04	0.05	0.02	0.57	-0.13	0.01	0.46	0.04	0.22
s'	a	distressing	travesty	that	offers	food	for	thought	.	
$A(s')$			Failed Prediction-Filter							-
s'	a	distressing	humorous	that	offers	food	for	thought	.	
$A(s')$			Failed POS-Filter							-

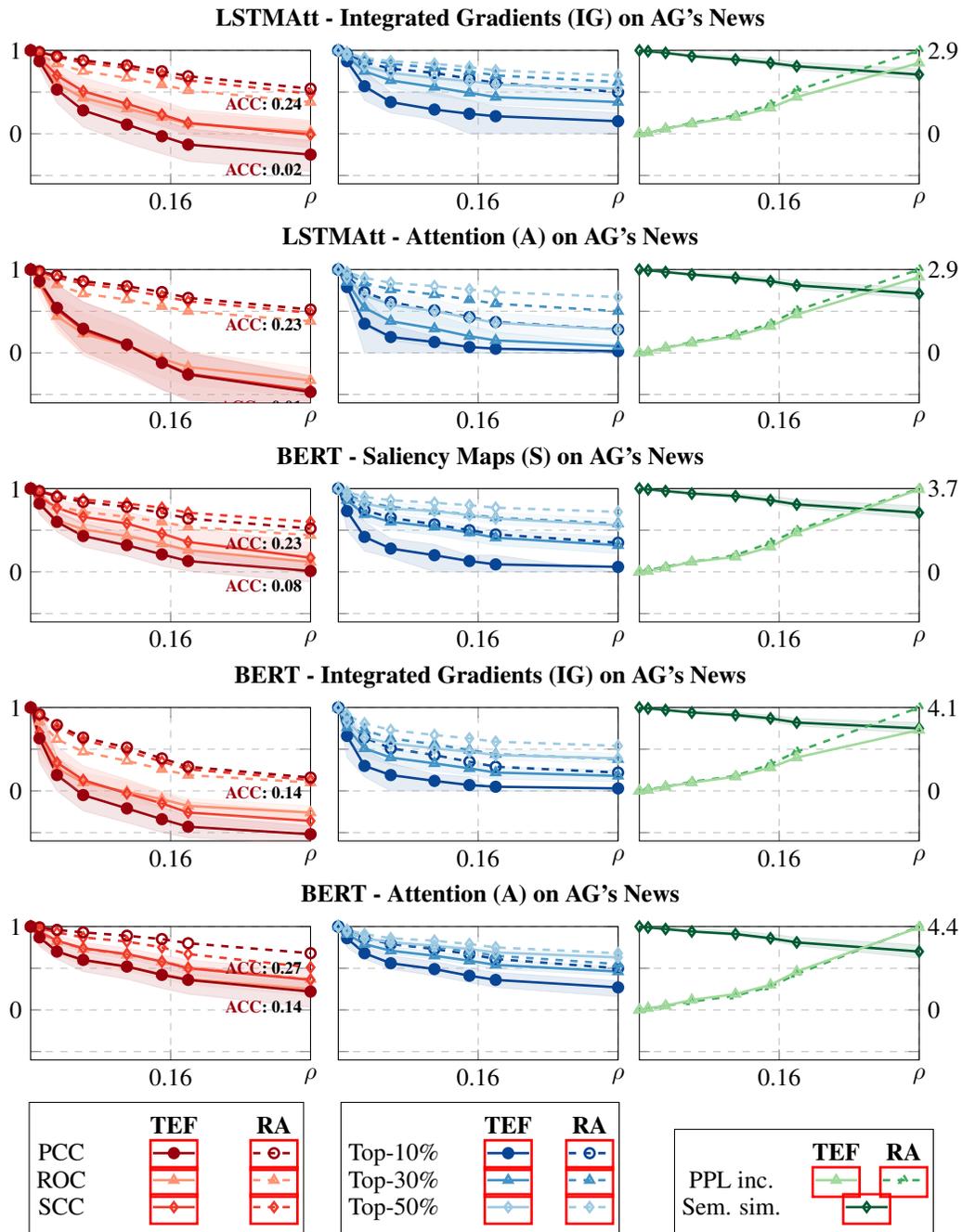
The **final adversarial sequence** s_{adv} becomes the valid s' with the lowest PCC value, which is given in the following table.

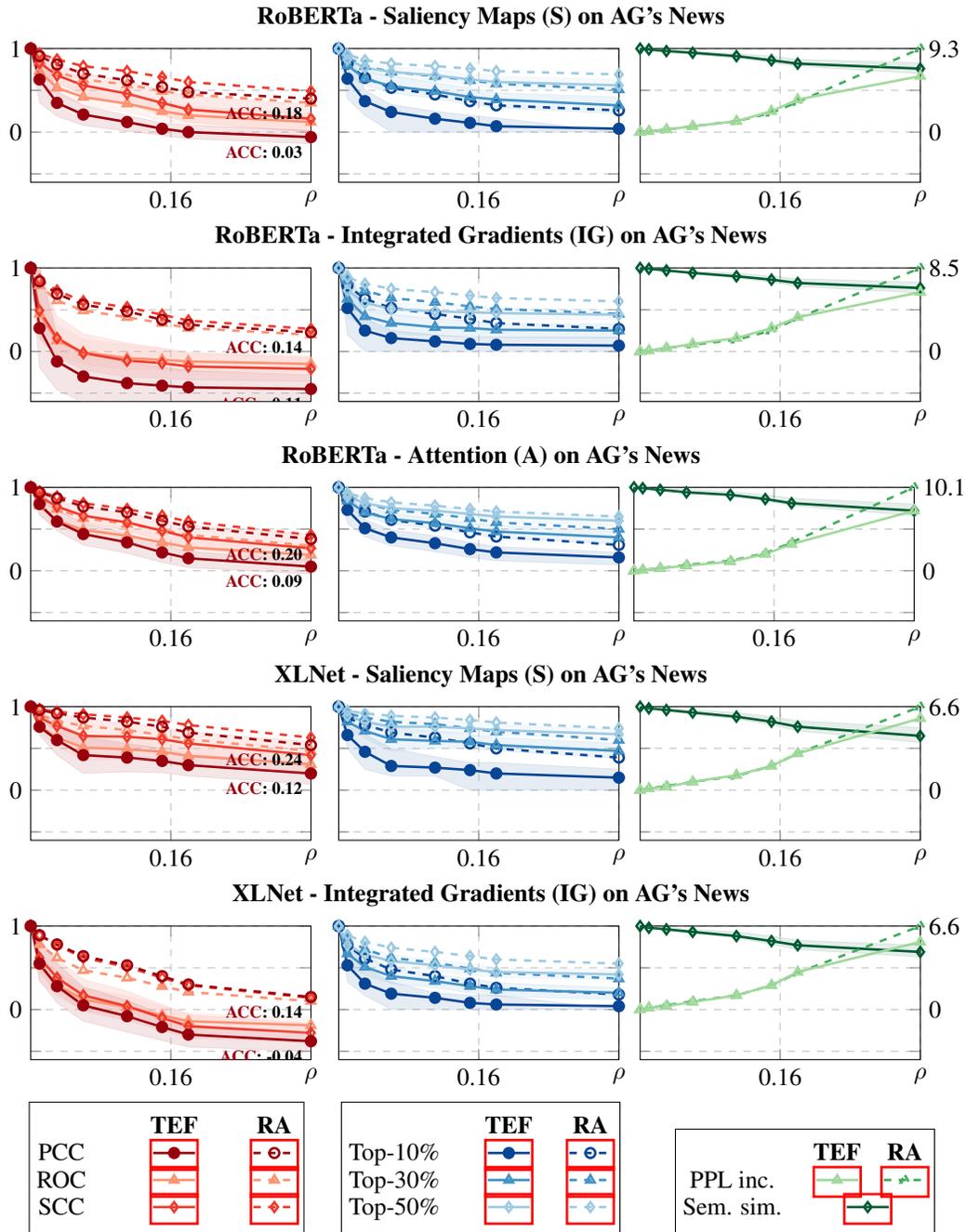
w_i	0	1	2	3	4	5	6	7	8	PCC
s	a	poignant	comedy	that	offers	food	for	thought	.	
$A(s)$	0.0	-0.08	0.4	0.05	0.16	0.23	0.03	0.22	0.0	-
s_{adv}	a	distressing	comic	that	offers	food	for	thought	.	
$A(s_{adv})$	-0.02	0.04	0.05	0.02	0.57	-0.13	0.01	0.46	0.04	0.22

A.2 ROBUSTNESS OF EXPLANATIONS

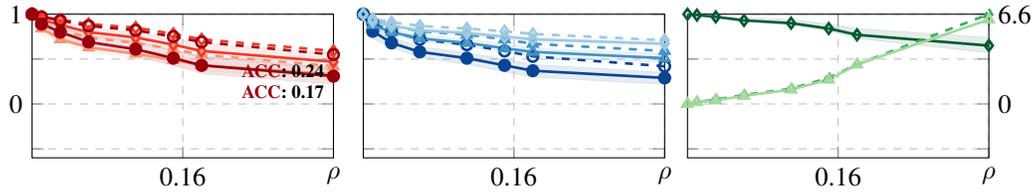
A.2.1 AG's NEWS





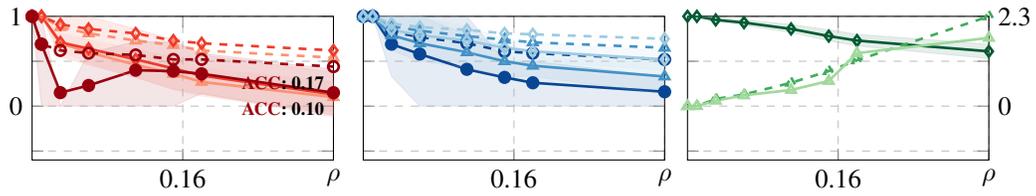


XLNet - Attention (A) on AG's News

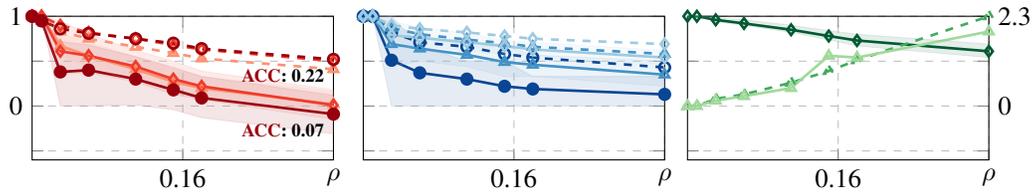


A.2.2 MR

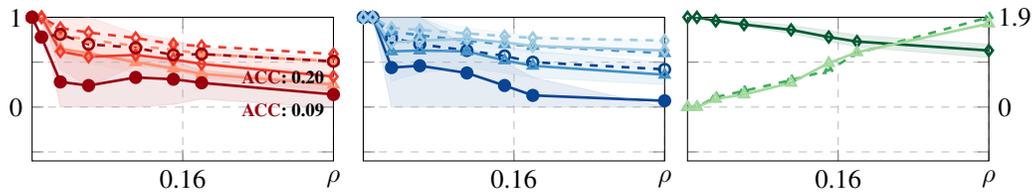
CNN - Saliency Maps (S) on MR



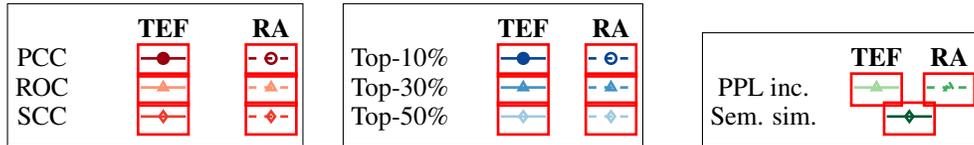
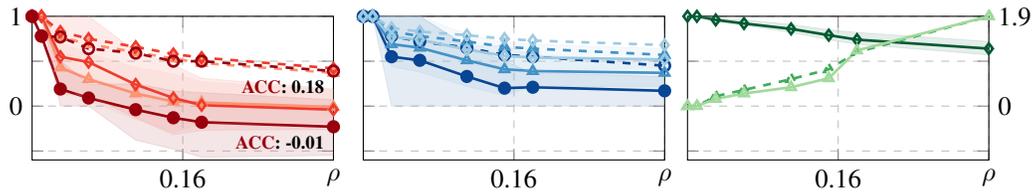
CNN - Integrated Gradients (IG) on MR

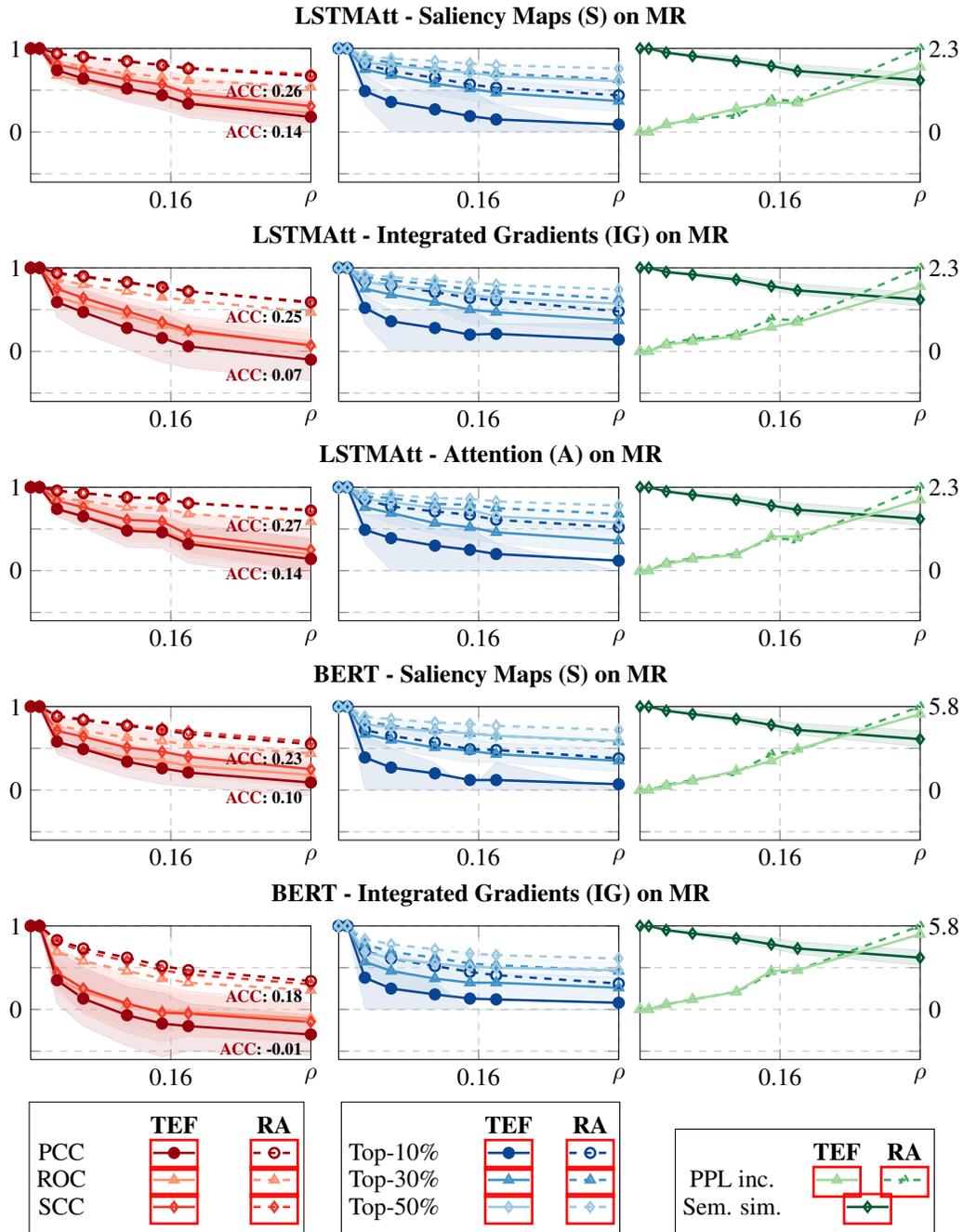


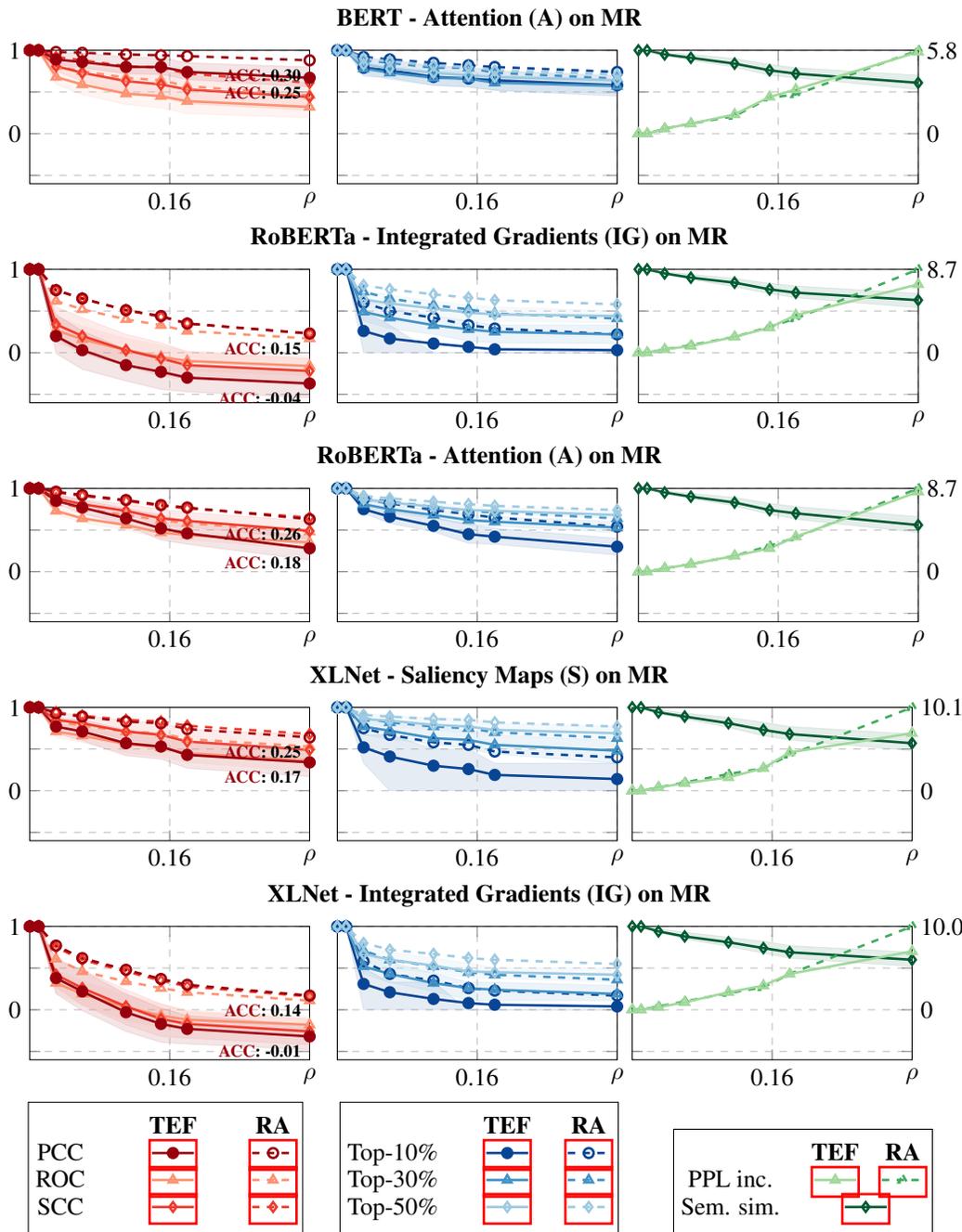
LSTM - Saliency Maps (S) on MR

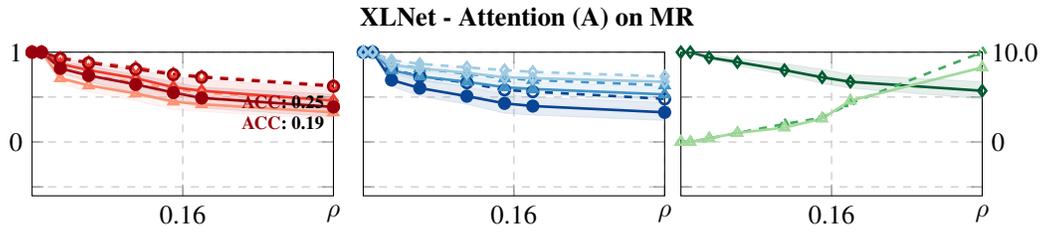


LSTM - Integrated Gradients (IG) on MR

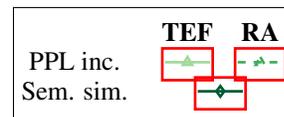
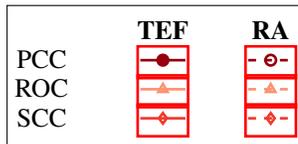
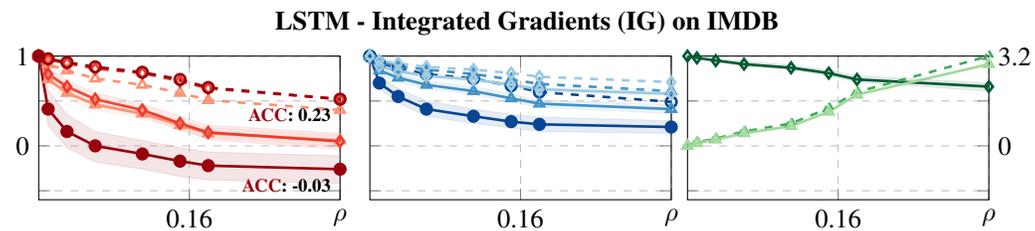
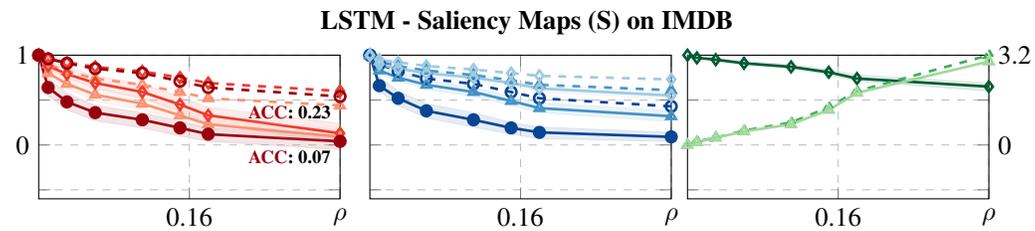
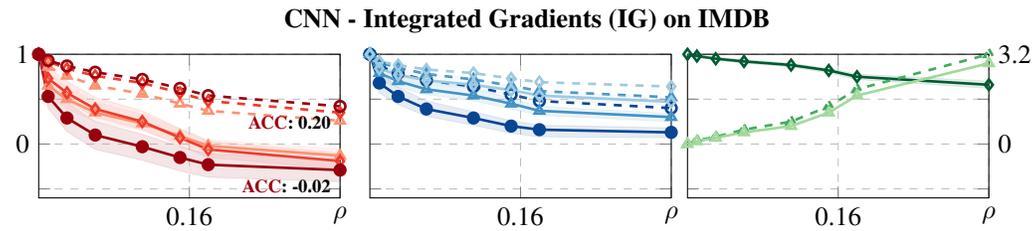
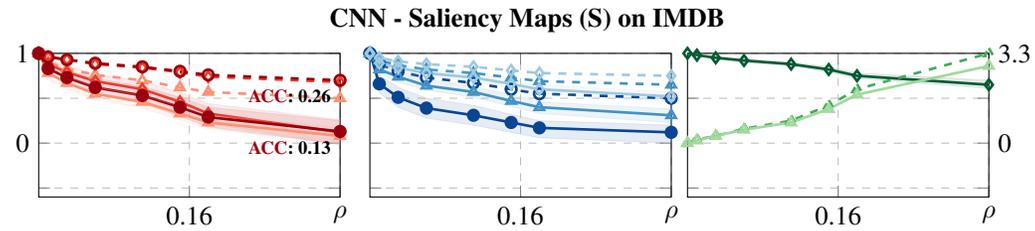


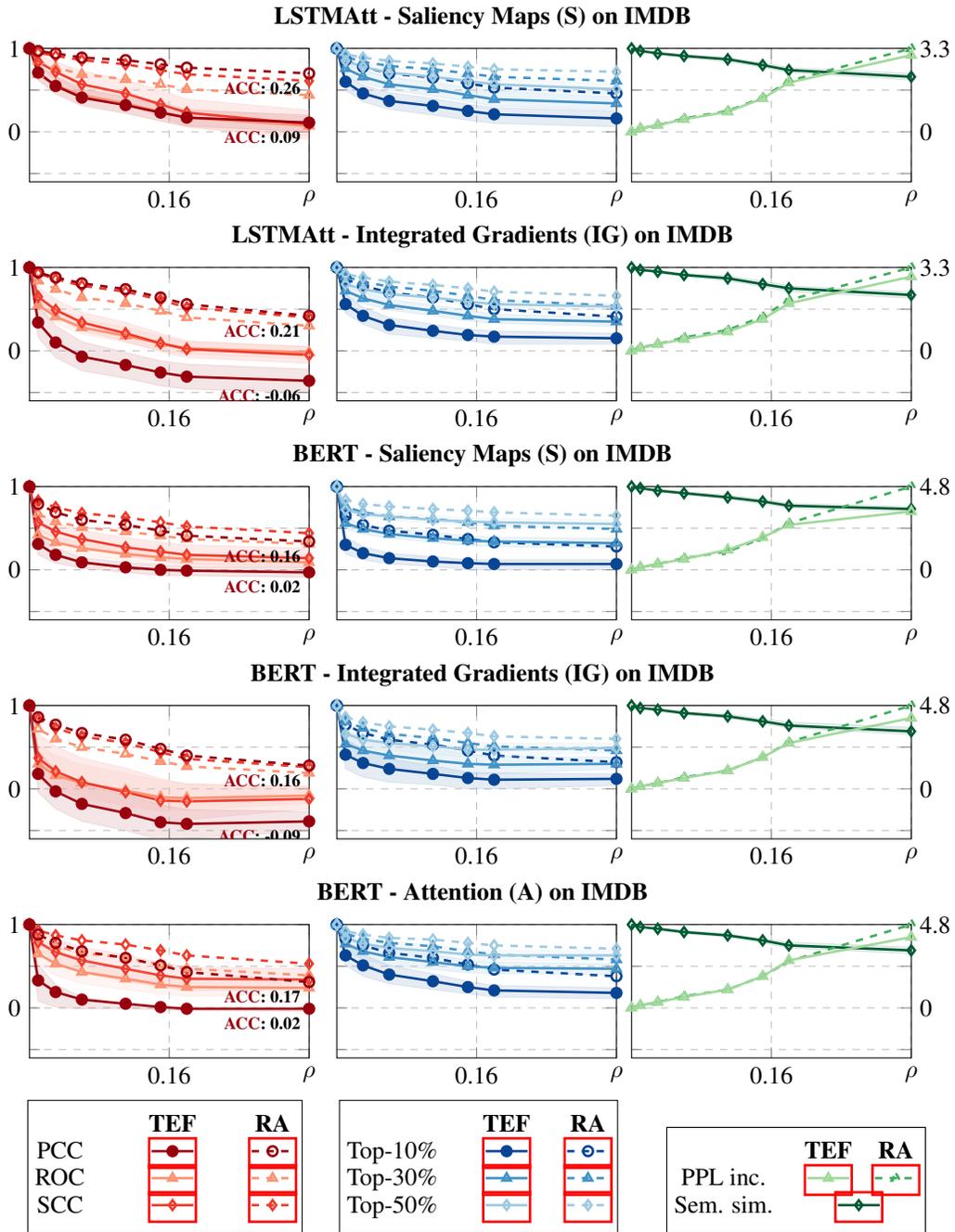


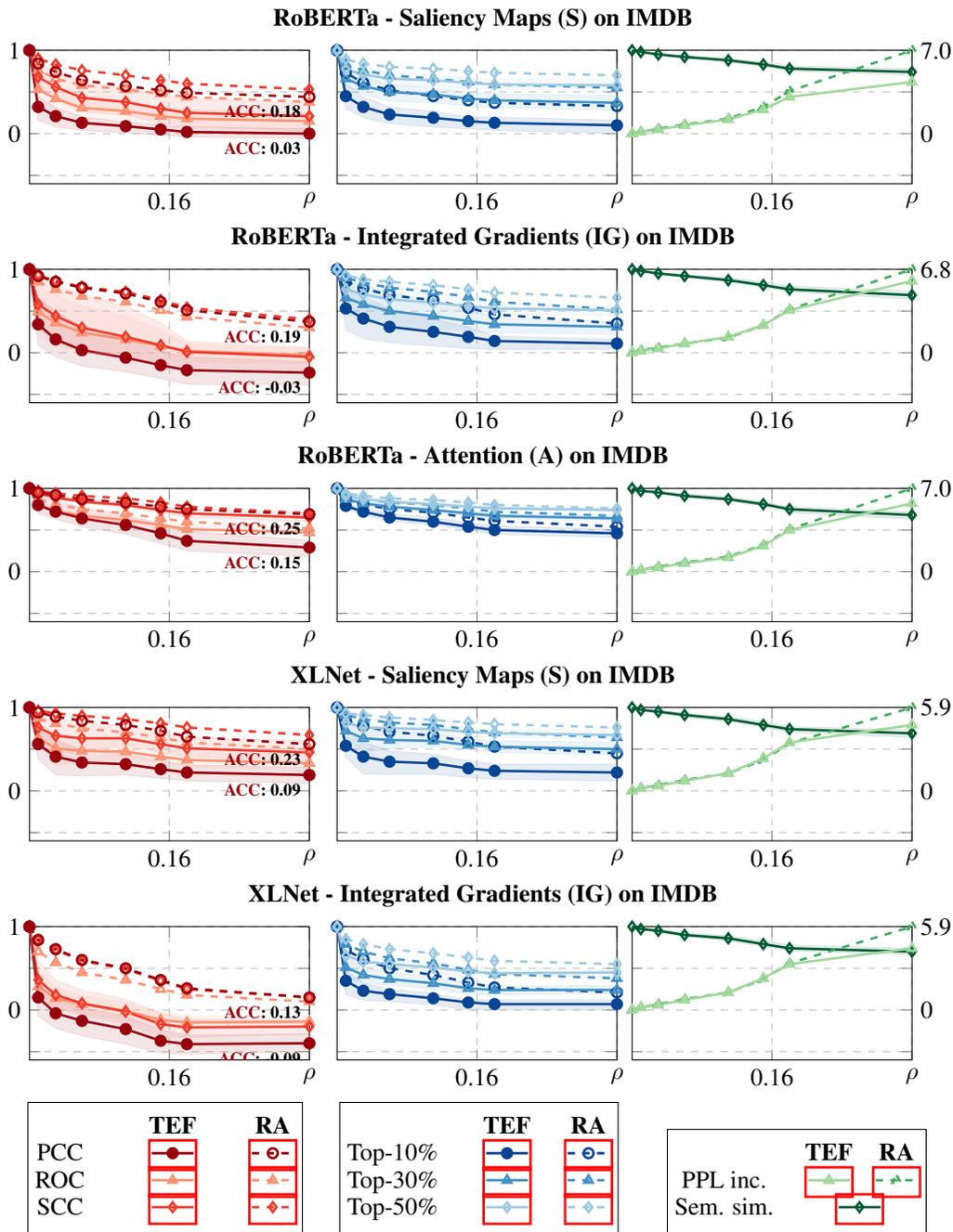




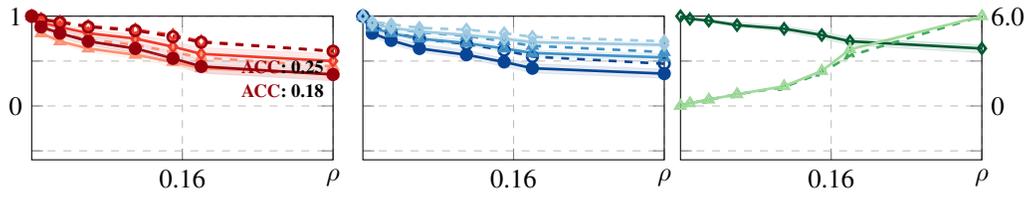
A.2.3 IMDB





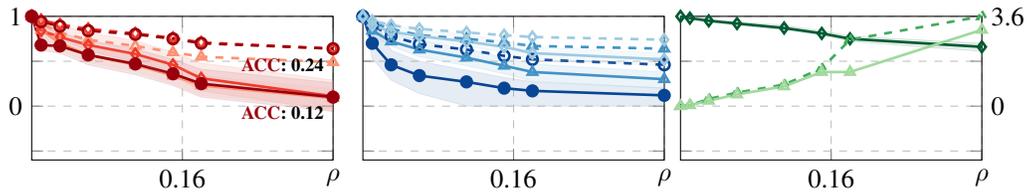


XLNet - Attention (A) on IMDB

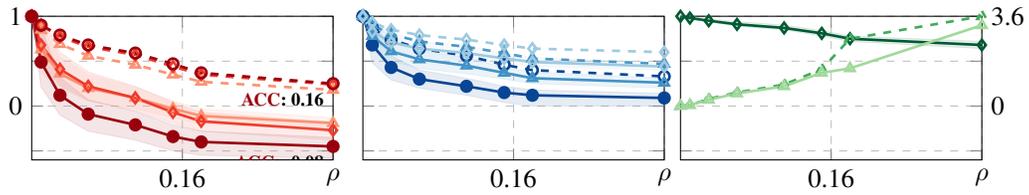


A.2.4 YELP

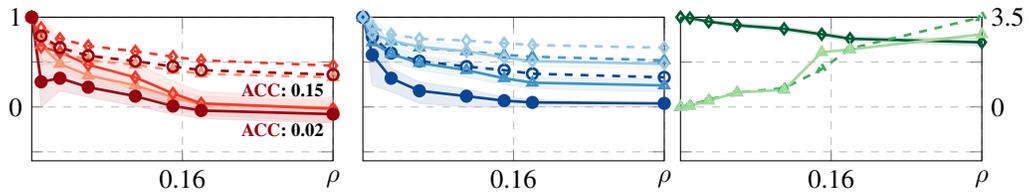
CNN - Saliency Maps (S) on Yelp



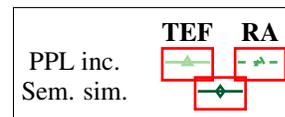
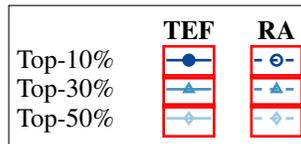
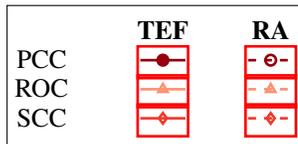
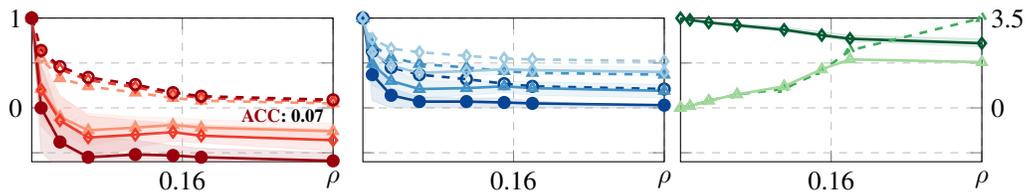
CNN - Integrated Gradients (IG) on Yelp

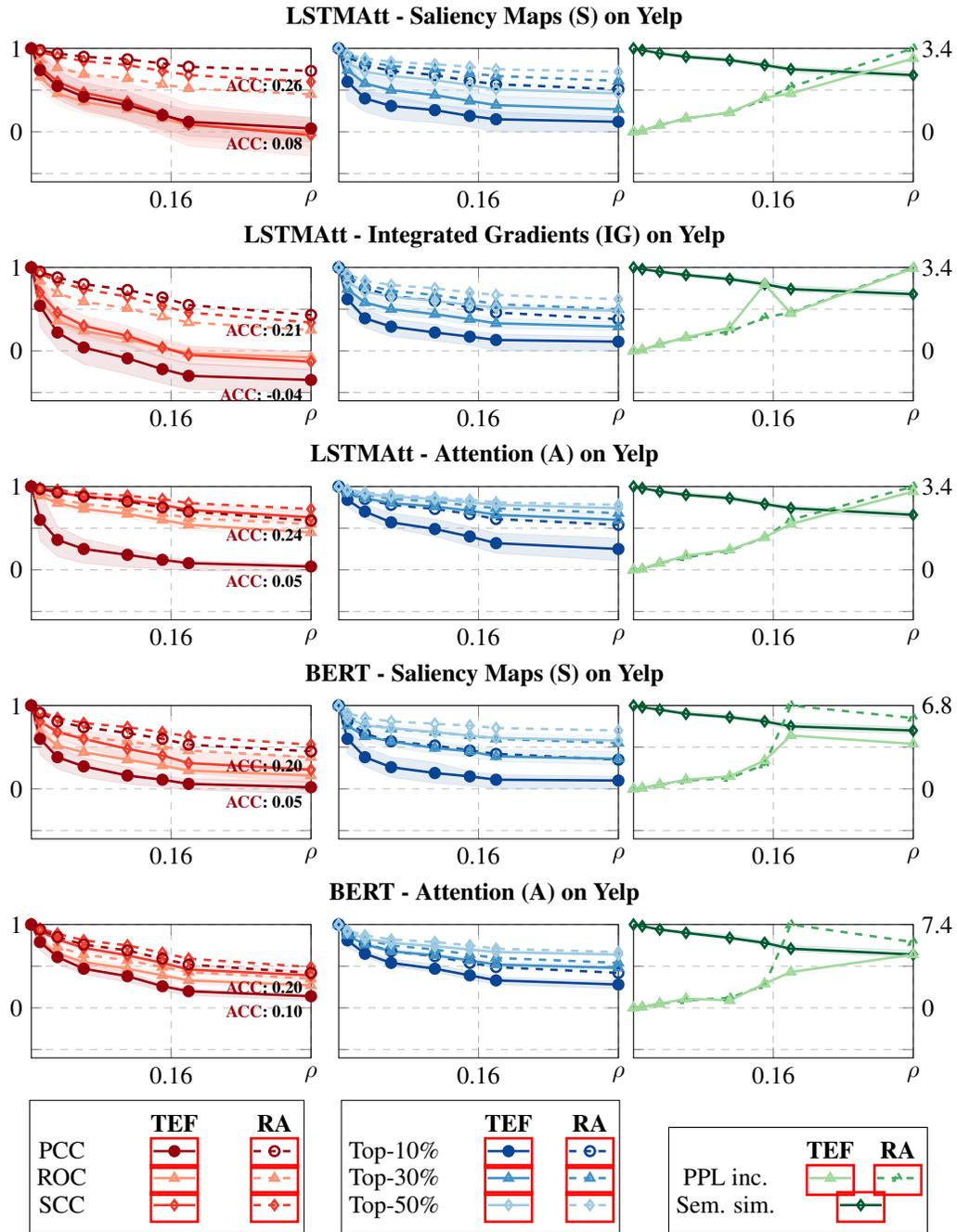


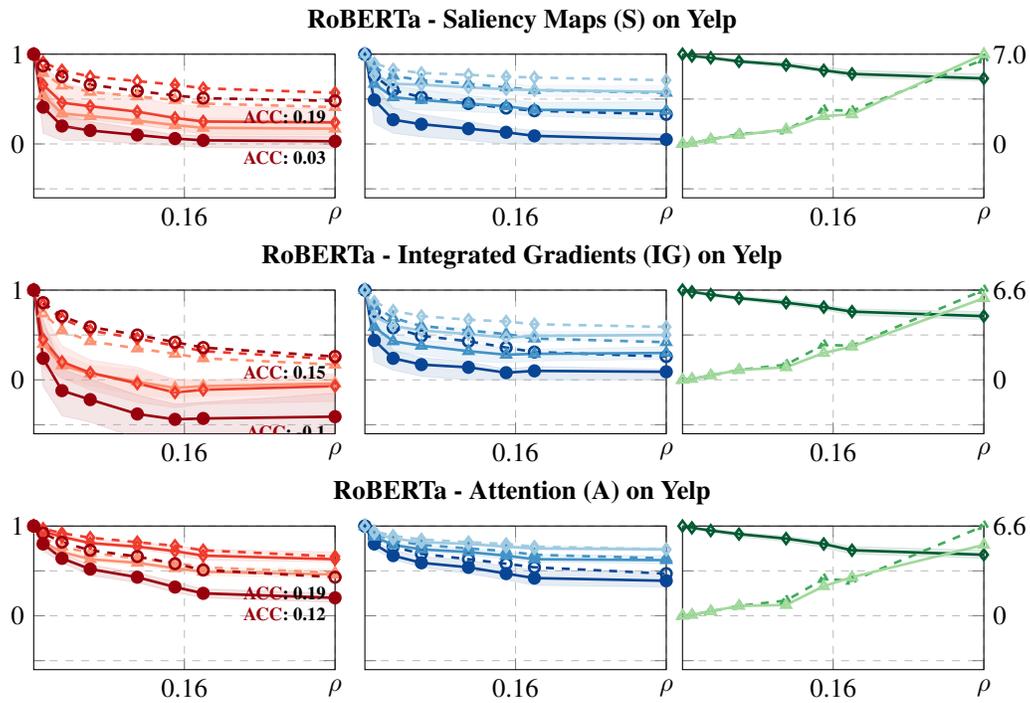
LSTM - Saliency Maps (S) on Yelp



LSTM - Integrated Gradients (IG) on Yelp







A.2.5 FAKE NEWS

