A Claim-only vs. Evidence-aware Classification

Table 5 shows the performance of the claim-only BERT classifier and numerous evidence-aware baseline in a three-class (SUPPORT, REFUTE, NOT ENOUGH INFO) setting.

Model	Accuracy
Majority Baseline	33.3
Evidence-Aware Classifiers	
DA	52.1
NSMN	69.7
Claim-Only Aware Classifiers	
InferSent (random emb.)	54.1
InferSent (GloVe)	57.3
BERT	61.7

Table 5: Results of evidence-aware and claim-only classifiers on the three label development set of the FEVER dataset.

B Additional Analysis

B.1 Fever Split

The split of the public FEVER dataset is described in Table 6.

Split	SUPPORT	REFUTE	NOT ENOUGH INFO
Training	80,035	29,775	35,639
Development	6,666	6,666	6,666
Total	86,701	36,441	42,305

Table 6: Fever dataset split.

B.2 Top LMI-ranked Bigrams in Train and Development Set

Table 7 and Table 8 summarize the top 10 bigrams for SUPPORT and NOT ENOUGH INFO. The correlation between the biased phrases in the two dataset splits is not as strong as in the REFUTE label, presented in the paper. However, one can notice that some of the biased bigrams in the training set, such as "least one" and "starred movie", translate to cues that can help in predictions over the development set. Bigrams are chosen for this exploratory analysis as they yield more comprehensible phrases.

	Train		Development	
Bigram	**LMI** $\cdot 10^{-6}$	$p(l\|w)$	**LMI** $\cdot 10^{-6}$	$p(l\|w)$
united states	271	0.64	268	0.44
least one	269	0.90	267	0.77
at least	256	0.72	163	0.48
person who	162	0.90	135	0.61
stars actor	143	0.86	111	0.71
won award	133	0.80	50	0.56
american actor	126	0.79	55	0.45
starred movie	100	0.88	34	0.80
from united	100	0.82	108	0.67
from america	96	0.89	108	0.74

Table 7: Top 10 LMI-ranked bigrams in the train set of FEVER for SUPPORT.

	Train		Development	
Bigram	**LMI** $\cdot 10^{-6}$	$p(l\|w)$	**LMI** $\cdot 10^{-6}$	$p(l\|w)$
worked with	221	0.40	129	0.56
s name	99	0.59	106	0.65
award winning	98	0.52	208	0.79
wyatt earp	96	0.42	*	0.00
finished college	86	0.68	10	0.42
and it	86	0.42	254	0.73
will ferrell	79	0.46	*	0.00
can be	75	0.35	72	0.48
and he	74	0.38	52	0.59
tim rice	70	0.41	*	0.00

Table 8: Top 10 LMI-ranked bigrams in the train set of FEVER for NOT ENOUGH INFO. * denotes computationally infeasible, as occurrence is zero in the development set.

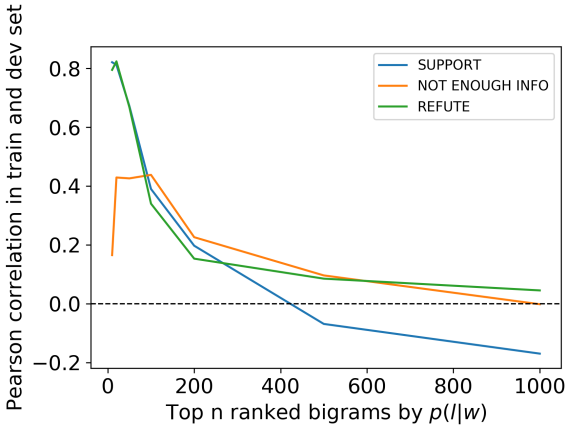

Figure 2: Pearson r scores of $p(l|w)$ for the top LMI-ranked bigrams in the train and development sets.

We calculated the Pearson correlation score between $p(l|w)$ for both train and development set. Figure 2 shows curves that start from very high correlations (i.e. 0.8 to 0.5) for the top $p(l|w)$-ranked ∼50-100 bigrams of REFUTE and SUPPORT (the curve for NOT ENOUGH INFO is less stable), dropping at around rank 400, supporting

the existence of 'give-away-bigrams' and that they are common in both training and development set.

B.3 Top Bigram Distribution in the Development Claims

Figure 3 illustrates the distribution of the top 1,000 LMI-ranked training set bigrams in the development set. In the case of the REFUTE class, we see that 57.6% of the REFUTE claims in the development set contain the top 1,000 LMI-ranked bigrams. Out of them, a high 59.5% are indeed labeled REFUTE. This concludes that 34.3% of all REFUTE claims are potentially biased. Following the same line of explanation, 32.8% and 16.2% of the SUPPORT and NOT ENOUGH INFO claims also face this problem.

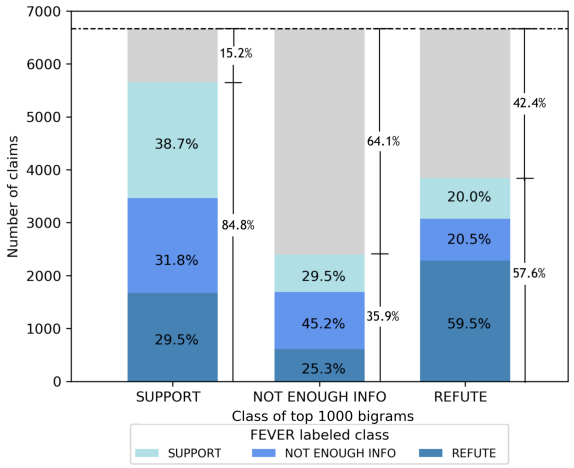

Figure 3: Percentage of claims containing at least one of the top 1,000 LMI-ranked bigrams (colors are used to express the class the claims were associated to). The overall heights of the bars indicate the number of claims expected for each class (i.e. 6,666).