



Figure 1: Number of words, POS tags, punctuation and numbers per statement in real and fake news in LIAR and COVID-19, and number of https-links per statement in COVID-19. The mean values are shown as white-filled circles in the plot.

the stylistic differences between real and fake news in the datasets. We use spaCy (<https://spacy.io>) to parse the statements and get the Part-of-Speech (POS) tags. For LIAR, we group “pants-fire”, “false”, and “barely-true” as fake and “half-true”, “mostly true”, and “true” as real. We compare the distribution of different words, POS tags (NOUN, PROPN, VERB, ADJ, ADV), punctuation, and number-like words in each statement in Figure 1.

The length of posts is quite different between the two classes in COVID-19, with an average of 32 and 22 for real and fake statements, respectively, as shown in Figure 1a. LIAR, on the other hand, has a similar statement length, with 18 words per statement for real and 17 for fake.

In general, COVID-19 has distinct linguistic features between classes whereas LIAR shows more similar features. In particular, COVID-19 contains links, mostly https links, which are listed as a separate category in Figure 1c, showing a very skewed distribution.

## Experiments and Results

For our experiments, we use ERNIE, three pre-trained KnowBert models with different KBs (Wiki, WordNet, W+W), KEPLER, and K-ADAPTER with three adapters (F, L, F-L) in the published implementation, fine-tune the models to our task and compare the result with the baseline models - BERT-base, RoBERTa-base, and RoBERTa-large.

**Detection Accuracy** The detection accuracy of the knowledge-enhanced PLMs and the corresponding baselines is shown in Table 1. On LIAR, all knowledge-enhanced methods improve over the baseline with KnowBert-W+W reaching the best overall result (improvement of +2.59 over BERT-base), whereas, on COVID-19, only three of eight models show improvement, and only by a small margin.

The computational cost varies per approach. KEPLER retains the baseline PLM architecture, thus there is no overhead compared to RoBERTa-base. K-ADAPTER also freezes the RoBERTa-large layers, but there is an overhead of 9-23% from the adapters, while the overhead for KnowBert is 40-87% and 111-131% for ERNIE.

MODEL	BASE	LIAR	COVID-19
<b>BERT-Base (BB)</b>	-	26.36 $\pm 0.58$	97.51 $\pm 0.19$
<b>RoBERTa-Base (RB)</b>	-	26.71 $\pm 0.93$	97.61 $\pm 0.26$
<b>RoBERTa-Large (RL)</b>	-	<b>27.36</b> $\pm 0.79$	<b>97.92</b> $\pm 0.17$
ERNIE	BB	27.53 $\pm 0.13$	97.30 $\pm 0.18$
KnowBert-Wiki	BB	27.64 $\pm 0.09$	97.37 $\pm 0.09$
KEPLER	RB	26.77 $\pm 1.15$	97.58 $\pm 0.15$
K-ADAPTER-F	RL	<b>28.63</b> $\pm 0.90^*$	<b>97.92</b> $\pm 0.10$
KnowBert-WordNet	BB	26.95 $\pm 0.45$	97.00 $\pm 0.06$
KnowBert-W+W	BB	<b>28.95</b> $\pm 0.64^*$	97.56 $\pm 0.15$
K-ADAPTER-L	RL	28.46 $\pm 0.87^*$	98.07 $\pm 0.09$
K-ADAPTER-F-L	RL	27.45 $\pm 0.78$	<b>98.11</b> $\pm 0.14$

Table 1: Detection accuracy results (average of five runs). The first section corresponds to the baseline models. Models in the second section use Wikidata KB. The third section shows models using other KBs and features. The best values within each section per dataset are marked in bold. The subscript numbers with  $\pm$  show the standard deviation. Results with \* indicate statistically significant improvements over the baseline, both for the mean (t-test, one-sided,  $p < .05$ ) and median (Wilcoxon signed rank test, one-sided,  $p < .05$ ).

**KB Linking** ERNIE and KnowBert create links between the text and KB entities at runtime and the quality of this linking influences the output. ERNIE uses TAGME and selects only one entity candidate per text span. In Figure 2 we show the 50 most frequently selected KB entities for each dataset. We can see that in COVID-19, the most frequent entities are not content-related (“https”, “twitter”) while “COVID-19”, the most frequent relevant term in the dataset, is missing in the linked entities. For LIAR, on the other hand, the linked entities seem relevant. Since LIAR was collected three years earlier, it is apparently a better match for the entity linker and the KB used. Another potential influence on the effectiveness of KB integration is the number of linked entities. In contrast to ERNIE, KnowBert selects the 30 most probable entities per text span. In a sensitivity study, we restrict KnowBert-W+W to only one entity, which reduces the accuracy on LIAR from 28.95% to

