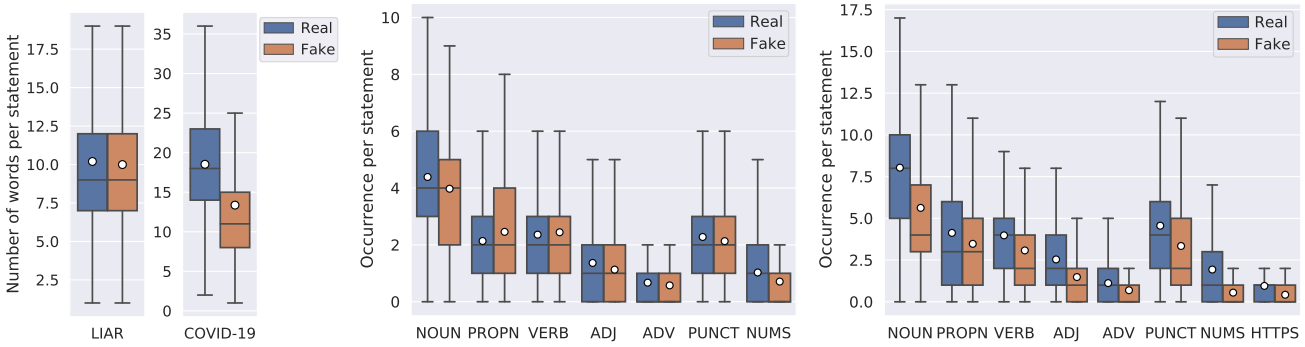

(a) Word count per statement (b) POS, punctuation, numbers in LIAR (c) POS, punctuation, numbers, https in COVID-19

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
BERT-**B**ase (BB)	-	26.36 $_{\pm 0.58}$	97.51 $_{\pm 0.19}$
RoBERTa-**B**ase (RB)	-	26.71 $_{\pm 0.93}$	97.61 $_{\pm 0.26}$
RoBERTa-**L**arge (RL)	-	**27.36** $_{\pm 0.79}$	**97.92** $_{\pm 0.17}$
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	**28.63** $_{\pm 0.90}^{*}$	**97.92** $_{\pm 0.10}$
KnowBert-WordNet	BB	26.95 $_{\pm 0.45}$	97.00 $_{\pm 0.06}$
KnowBert-W+W	BB	**28.95** $_{\pm 0.64}^{*}$	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 $_{\pm 0.78}$	**98.11** $_{\pm 0.14}$

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

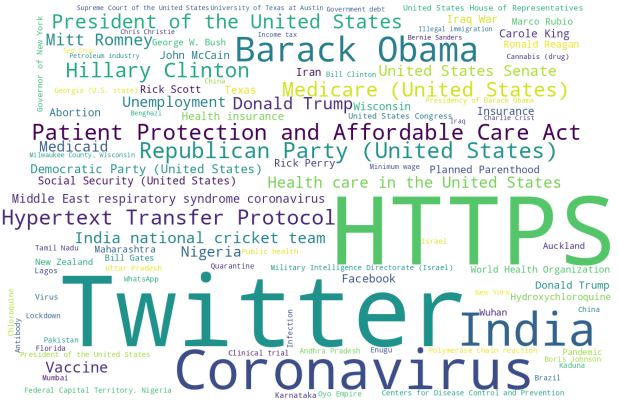

Figure 2: Word clouds for the 50 most frequent entities linked by ERNIE in LIAR (top) and COVID-19 (bottom).

27.31%, below the accuracy of ERNIE (27.53%).

Discussion

The reliable improvement of detection accuracy on LIAR by integrating PLMs with Wikidata shows the potential of knowledge integration exceeding the results obtained by integrating multiple types of metadata by **?**. On the other hand, the improvements are good but not dramatic for LIAR and not consistent for COVID-19. We can identify two aspects contributing to the result which are relevant to the effective use of knowledge-enhanced models:

1) Currentness and relevance of the KB: as COVID-19 was collected after most of the PLMs were trained, some terms such as "COVID-19" are not in the KB;

2) Quality of the dataset: the COVID-19 dataset contains confounders that provide strong cues, overshadowing the impact of the knowledge base. The most important one is the occurrence of https links, which appear in 95.3% of the real posts but only 42.3% of the fake posts.

There is also potential to achieve more explainability and interpretability with direct KB integration at runtime. Take this statement from COVID-19: *"DNA Vaccine: injecting genetic material into the host so that host cells create proteins that are similar to those in the virus against which the host then creates antibodies"* as an example, KnowBert-W+W correctly classifies it as "real", whereas BERT-base fails. We observe most mention spans in the statement, i.e. *"DNA"*, *"injecting"*, *"genetic"*, *"genetic material"*, *"host"*, *"cells"*, etc. are correctly linked to entities *"DNA"*, *"Injection_(medicine)"*, *"Genetics"*, *"Genome"*, *"Host_(biology)"*, *"Cell_(biology)"*, respectively, therefore it seems that the entity links may have contributed to KnowBert-W-W for this classification. However, the level of explainability is still limited.

Application Aspects Automatic fake news detection in practice adds two dynamic application aspects, which are difficult to test with static datasets as our experiment on COVID-19 has shown:

(1) Dynamic adaptation: it is necessary to update the system to the changing characteristics of real and fake news (**?**). Knowledge-enhanced models that use KBs at runtime offer an opportunity to update the KB independent of the model. This has the advantage that fake news can be recognised as contradicting the KB before there are any fake news examples.

(2) Adversarial robustness: fake news authors are very likely to take evasive action. Adapting the text style is relatively easy and could be automated, which makes the detection with stylistic features difficult (see **??**).

Deployment of fake news detection in social media will also need human verification, e.g. when a user challenges actions taken against them. Here, KB integration can offer the advantage of insight into knowledge that has been used in the detection for better explainability.

Related Work

In recent years large-scale PLMs i.e. BERT and RoBERTa have dominated NLP tasks, including some content-based fake news detection (**?**). Most fake news detection approaches either combine text with metadata (e.g. **?**) or focus only on the source of the text (e.g. **?**). For LIAR, **?** extend the data with evidence sentences in a new dataset called LIAR-PLUS to improve detection. **?** introduce a Deep Averaging Network to model the discursive structure of the text and use Siamese models on the extended text data. **?** predict labels at two levels of granularity. For COVID-19, there are a number of results from the CONSTRAINTS 2021 workshop (**?**) which use a wide variety of traditional and neural NLP models. None of these approaches uses external knowledge, so they could all benefit from KB integration.

Conclusion and Future Work

In this paper, we study the effectiveness of enhancing PLMs with knowledge bases for fake news detection. We find that integrating knowledge with PLMs can be beneficial on a static dataset but it depends on suitable KBs and the quality of the data. On the modelling level, there are many routes for improvement. For practical application, more insight into what knowledge is used would be useful as well as dynamic adaptation of the models and the KBs. Integrating KBs with PLMs offers potentially more robust and timely fake news detection. However, a new evaluation approach, i.e. a testing scenario that models dynamic knowledge as well as adversarial and automatic fake news generators, is needed to assess the true potential of knowledge integration.

Totam voluptatibus accusamus harum facere expedita ut, minus accusantium aliquam distinctio optio illo voluptatem fugit sint cupiditate, laborum explicabo culpa itaque incidunt molestias iure error necessitatibus repudiandae, aperiam distinctio ipsum?Blanditiis doloremque numquam sequi accusamus facere, dignissimos ipsum itaque obcaecati minima dolor autem quibusdam voluptas odit assumenda quod, delectus sunt velit error quibusdam eius nostrum quis enim necessitatibus, soluta vitae illum, nulla repellat dolore omnis modi voluptas veritatis maiores rerum aperiam?Asperiores expedita ducimus enim, tempore quasi culpa ratione eius laboriosam quis facere nam, aliquid quos sequi assumenda