Evaluation of Fake News Detection with Knowledge-Enhanced Language Models

Chenxi Whitehouse, Tillman Weyde, Pranava Madhyastha, Nikos Komninos

City, University of London
{chenxi.whitehouse, t.e.weyde, pranava.madhyastha, nikos.komninos.1}@city.ac.uk

Abstract

Recent advances in fake news detection have exploited the success of large-scale pre-trained language models (PLMs). The predominant state-of-the-art approaches are based on fine-tuning PLMs on labelled fake news datasets. However, large-scale PLMs are generally not trained on structured factual data and hence may not possess priors that are grounded in factually accurate knowledge. The use of existing knowledge bases (KBs) with rich human-curated factual information has thus the potential to make fake news detection more effective and robust. In this paper, we investigate the impact of knowledge integration into PLMs for fake news detection. We study several state-of-the-art approaches for knowledge integration, mostly using Wikidata as KB, on two popular fake news datasets - `LIAR`, a politics-based dataset, and `COVID-19`, a dataset of messages posted on social media relating to the COVID-19 pandemic. Our experiments show that knowledge-enhanced models can significantly improve fake news detection on `LIAR` where the KB is relevant and up-to-date. The mixed results on `COVID-19` highlight the reliance on stylistic features and the importance of domain-specific and current KBs. The code is available at https://github.com/chenxwh/fake-news-detection.

Introduction

The world is witnessing a growing epidemic of fake news, which includes misinformation, disinformation, rumours, hoaxes, and other forms of rapid spread and factually inaccurate information (**?**). Fake news has been observed to severely impact political processes because of the wide reach of social media (**?**). Misinformation related to medical issues, such as the COVID-19 pandemic, can cost lives (**?**). Automated methods for fake news detection and mitigation are a critical yet technically challenging problem (**?**).

In this paper, we focus on content-based fake news detection: methods that assess the truthfulness of news items based only on the text without using metadata. State-of-the-art models for this task are driven by advances in large-scale pre-trained language models (PLMs) (e.g. **??**), which are trained on vast amounts of raw web-based text using self-supervised methods (**?**). A major limitation of these models is the lack of explicit grounding to real-world entities and relations, which makes it difficult to recover factual knowledge (**?**). On the other hand, knowledge bases (KBs) provide a rich source of structured and human-curated factual knowledge, often complementary to what is found in raw text. This has recently led to the development of KB-augmented language models. Fake news detection can particularly benefit from the integration of KBs, making such models less dependent and reliant on surface-level linguistic features.

In this study, we empirically analyse the impact of recent state-of-the-art knowledge integration methods, which enhance PLMs with KBs, for content-based fake news detection tasks. We evaluate ERNIE (**?**), KnowBert (**?**), KEPLER (**?**) and K-ADAPTER (**?**) on two distinct publicly available datasets: `LIAR` (**?**), a politically oriented dataset, and `COVID-19` (**?**), a dataset related to the recent pandemic. We find that integrating knowledge can improve fake news detection accuracy, given that the knowledge bases are relevant and up-to-date. Our experiments are not designed to find new state-of-the-art models for these datasets, but to investigate the effect of knowledge base integration into PLMs.

Our contributions are as follows: we evaluate multiple KB integration methods for fake news detection, we investigate model and data aspects that can prevent KB integration from being effective or from being effectively measured, and we discuss the potential for real-world applications.

In the following sections, we present a brief overview of four state-of-the-art methods that integrate KBs with PLMs studied in this paper. We then introduce and compare the datasets, the experiments with different knowledge-enhanced models, and the effectiveness of entity linking. We discuss our findings with respect to the necessary conditions for KB integration to be effective and how to assess its effect in application scenarios. Finally, we discuss the challenges in fake news detection and promising future directions.

Method

In this section, we introduce the models with KB integration and describe the datasets and our experimental setup.

Knowledge Integration for PLM

Standard deep learning models obtain information from predicting and classifying text as they are trained, but have no prior knowledge of, or interaction with, world knowledge.

Although PLMs can effectively characterise linguistic patterns from text to generate high-quality context-aware representations, they are limited in their grasp of knowledge, concepts, and relations, which are essential for some Natural Language Processing (NLP) tasks, including assessing the truthfulness of news items.

On the other hand, KBs like Wikidata (https://www.wikidata.org) and WordNet (**?**) contain rich curated information about the world. Thus, they could greatly complement PLMs if effective integration methods were available. Several efforts have been made to integrate KBs into PLMs. In this paper, we study the following models:

ERNIE injects knowledge into BERT (**?**) by pre-training a language model on both large corpora and KBs. It uses TAGME (**?**) to link entities to Wikidata. TAMGE identifies entity mentions in the input text and links them to associated entity embeddings, which are then fused into the corresponding positions of the text. The knowledge-based learning objective is to predict the correct token-entity alignment. ERNIE has enhanced performance over BERT in entity typing and relation classification (**?**).

KnowBert incorporates KBs into BERT using a knowledge attention and re-contextualisation mechanism. It identifies mention spans in the input text and incorporates an integrated entity linker to retrieve entity embeddings from a KB. The entity linker is responsible for entity disambiguation, which considers 30 entity candidates and uses their weighted average embedding. Knowledge-enhanced entity-span representations are then re-contextualised with a word-to-entity attention technique. KnowBert has shown improvement over BERT in relationship extraction, entity typing and word sense disambiguation (**?**).

KEPLER integrates factual knowledge into PLMs by adding a knowledge embedding objective with the supervision from a KB and optimising it jointly with language modelling objectives. KEPLER is trained to encode the entities from their contextual descriptions, which enhances the ability of PLMs to extract knowledge from text. By keeping the original structures of PLMs, KEPLER can be used in general downstream NLP tasks without additional inference overhead. It is shown that KEPLER improves performance over RoBERTa (**?**) in relationship extraction, entity typing and link prediction (**?**).

K-ADAPTER retains the PLMs unchanged, but adds learnable adapter features that are trained in a multi-task setting on relation prediction and dependency-tree prediction. Two kinds of knowledge adapters have been developed by **?**: factual knowledge obtained from automatically aligned text triples on Wikipedia and Wikidata, and linguistic knowledge obtained via dependency parsing. Both have been found to improve relation classification, entity typing and question answering (**?**).

Datasets

In our experiments, we use `LIAR` and `COVID-19` to study fake news detection. They both consist of short statements, but with different content, time of collection, and linguistic and stylistic features.

`LIAR` was collected in 2017 from Politifact (https://www.politifact.com). It includes 12.8k human-labelled short statements about US politics from various contexts, i.e. news releases, TV interviews, campaign speeches, etc. Each statement has been rated for truthfulness by a Politifact editor using a six-grade scale: "pants-fire", "false", "barely-true", "half-true", "mostly true", and "true". `LIAR` also provides metadata (e.g. speaker, context), which we do not use in our experiments. While **?** has been widely cited, we only found three other results for our specific task (no metadata, six classes): (**???**), the latter has the current best accuracy of 34.5%.

`COVID-19` was collected in 2020 after the COVID-19 outbreak. It consists of 10.5k posts related to the pandemic which are obtained from different social media sites including Twitter, Facebook, and Instagram. The fake posts were collected from various fact-checking websites, i.e. Politifact and NewsChecker (https://newschecker.in), and the real posts were from Twitter using verified Twitter handles. Each post has a label, "real" or "fake". It was used as a shared task in the CONSTRAINT 2021 workshop (**?**) with the best-reported accuracy of 98.69%.

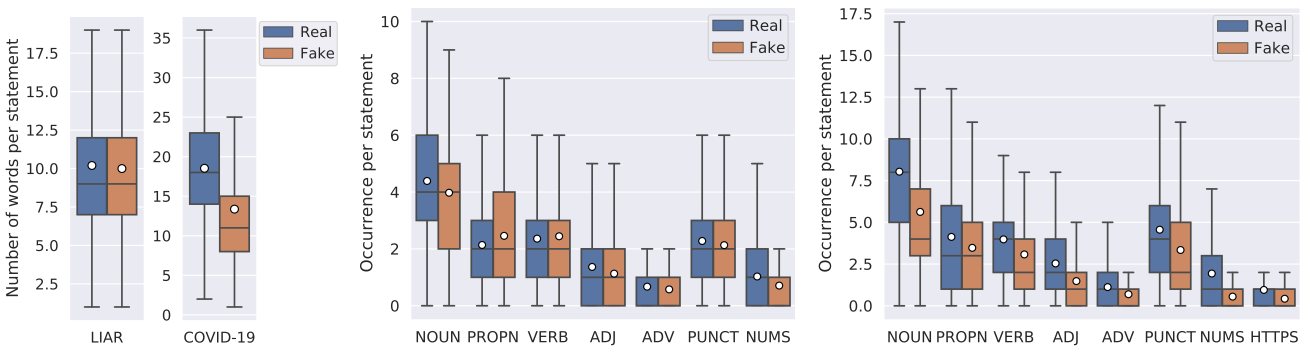
Experimental Setup

We use an empirical approach to study the effect of knowledge integration on fake news detection, to understand how knowledge is used by the model, and to evaluate the quality of the entity linker to the KB.

ERNIE and KnowBert are built on BERT-base, whereas KEPLER and K-ADAPTER are enhanced from RoBERTa-base and RoBERTa-large, respectively. We follow the concept of an ablation study to investigate the influence of external knowledge by comparing the performance of each knowledge-enhanced PLM with the corresponding baseline model. We note that ERNIE and KnowBert incorporate entity embeddings that are linked to the input. Therefore we visualise the entities linked that contribute to the fake news detection task in ERNIE, and design experiments to investigate the impact of entity disambiguation of KnowBert.

We evaluate the performance of the models on fake news detection by fine-tuning the knowledge-enhanced PLMs on the training set with the same hyperparameter settings. The input text is fed first to the PLM, and followed by a dropout ($p = 0.1$) and a linear layer. The output is then passed to a softmax layer for classification. We use AdamW optimiser (**?**) (learning rate of 5×10^{-6}) and cross entropy as the loss function. The maximum input length is set to 128, and the batch size is 4. We train for 10 epochs and usually observe convergence after five. We perform five runs for each experiment and report the average accuracy with the standard deviation. Both `LIAR` and `COVID-19` are already divided into train, validation, and test splits, which we use in our experiments as provided.

Linguistic Feature Analysis We also perform linguistic feature analysis following the work in **?** to investigate



(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 $_{\pm0.58}$	97.51 $_{\pm0.19}$
RoBERTa-**B**ase (RB)	-	26.71 $_{\pm0.93}$	97.61 $_{\pm0.26}$
RoBERTa-**L**arge (RL)	-	**27.36** $_{\pm0.79}$	**97.92** $_{\pm0.17}$
ERNIE	BB	27.53 $_{\pm0.13}$	97.30 $_{\pm0.18}$
KnowBert-Wiki	BB	27.64 $_{\pm0.09}$	97.37 $_{\pm0.09}$
KEPLER	RB	26.77 $_{\pm1.15}$	97.58 $_{\pm0.15}$
K-ADAPTER-F	RL	**28.63** $_{\pm0.90}^*$	**97.92** $_{\pm0.10}$
KnowBert-WordNet	BB	26.95 $_{\pm0.45}$	97.00 $_{\pm0.06}$
KnowBert-W+W	BB	**28.95** $_{\pm0.64}^*$	97.56 $_{\pm0.15}$
K-ADAPTER-L	RL	28.46 $_{\pm0.87}^*$	98.07 $_{\pm0.09}$
K-ADAPTER-F-L	RL	27.45 $_{\pm0.78}$	**98.11** $_{\pm0.14}$

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 ± 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

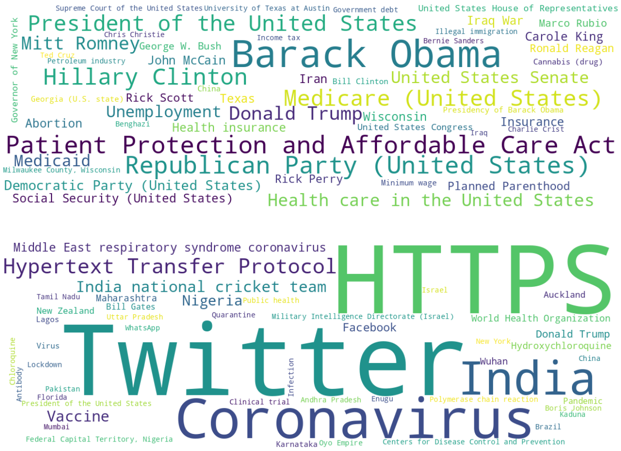


Figure 2: Word clouds for the 50 most frequent entities linked by ERNIE in LIAR (top) and COVID-19 (bottom).

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.

Obcaecati iste et labore, similique alias animi natus nisi quia voluptas assumenda, temporibus magni quos, aliquam accusantium ipsum, odio iusto nisi beatae.Ex iste accusamus laboriosam totam repellendus perspiciatis veniam cumque sunt est dolorem, sunt aliquid libero, facere sit veritatis recusandae eveniet perferendis doloribus animi.Repudiandae doloribus illo, incidunt voluptates eligendi sequi dicta architecto ab, placeat nisi aut modi doloremque pariatur magnam rem?Quaerat mollitia earum id praesentium et officiis, odio voluptatum nobis aliquam quam cumque numquam rerum temporibus, ullam non quidem quisquam neque nobis quia, velit aliquam ipsam suscipit officiis, iure neque perferendis?