What (Knowledge-Free) Fake News Classifiers Learn

Anonymous EMNLP-IJCNLP submission

Abstract

Unlike other types of document classifiers, fake news detectors are trained to detect signals in texts that their authors are trying to hide. In the absence of authoritative knowledge sources, detectors therefore need to rely on – at best – very subtle stylistic features. Through cross-corpora experiments, we show that even if such features exist, we are unlikely to learn them from modest-sized labelled datasets. Instead our current models learn undesired correlations that generalise poorly and bias our models in unfortunate ways.

1 Introduction

In recent years, we have witnessed a surge in research on detecting fake news, most of which amounts to training classifiers to distinguish between posts labelled fake and non-fake by professional annotators or mechanical turkers, based on surface patterns (Pérez-Rosas et al., 2018). Some of these classifiers are knowledge-free or non-grounded, i.e., they make their predictions in the absence of authoritative sources of information without knowing what the facts are; and hence, are strictly speaking models trying to detect what *sounds* fake rather than what *is* fake.

Fake news datasets are finite samples of moderate size, and state-of-the-art models are extremely high-dimensional. This means that fake news detection is subject to the curse of dimensionality (Bellman, 1957). In a context where the surface features we are looking for are – at best – subtle, we are very prone to over-fitting. We therefore hypothesise that fake news detectors generalise poorly across datasets and pick up on linguistic patterns that reflect the demographics of the authors who happened to produce the fake posts in our sample, rather than whether the posts are fake. If this is the case, fake news detectors are likely to be dangerously biased, being more likely to predict posts of certain demographics to be fake.

Contributions We present cross-sample experiments, across three fake news datasets, as well as thorough analyses thereof, to evaluate a standard approach to fake news detection. Having established poor cross-sample generalisation, we show over-fitting leads to undesired demographic bias.

	LSTM	bi-LSTM
LIAR (dev)	61.4	60.0
LIAR (test)	61.2	60.2
KAGGLE	44.3	45.4
FAKENEWSCORPUS	50.0	49.7

Table 1: Accuracy of LSTM models trained on LIAR, on binarised and balanced datasets. On out-of-sample data, performance drops to random or below.

	MRR	Kendall's τ	Spearman's ρ
LIAR_2	0.0003	$^*\tau$=0.0231	$^*\rho$= 0.0339
LIAR_2 (*Random*)	0.0006	τ=-0.0009	ρ=-0.0014
KAGGLE	0.0004	τ=-0.0003	ρ=-0.0006
FNC	0.0004	τ=0.0014	ρ=0.0019

Table 2: The similarity of feature importance rankings from linear models with those induced by a linear model trained on LIAR_1, according to three standard metrics. All models use a shared vocabulary. * denotes significance ($p < .05$). Note the correlation across different samples is as low as the correlation between the baseline and the random model.

2 Cross-Corpora Experiments

2.1 Datasets

The LIAR dataset[1] (Wang, 2017) is considered a benchmark dataset for the development of ap-

[1] https://www.cs.ucsb.edu/~william/data/liar_dataset.zip

LIAR$_1$	LIAR$_2$	KAGGLE	FAKENEWSCORPUS
is spending	million from	anti	going on many
this is the	there were	october	next by
holding	president barack obama has	non	article has been
few months	and all	year old	ps all rights reserved source
patient	arrested	2016	there something else
impose	education funding	co	with the help
of the population	texas public	self	and the day
spending in	muslim	november	of christ and
anybody	of business	share	bitcoin blockchain
country that	38	us	other words
says charlie	virginias	november 2016	this article has been
governments	raise taxes on	source	understand the difference
we dont have	debunked	via	lord in
per year	has already	print	in terms of
citizens who	in the entire	click	los angeles
of the states	he didnt	president elect	republished with permission from
was in	attorney general	us pro	the help of
tax hikes	extending	hillary	well as
might	in one	26	christ and the
over half	new taxes	so called	by lisa haven

Table 3: Top-20 most important features for the fake class in a linear model across four different datasets when using a shared vocabulary.

proaches to fake news detection and automatic fact-checking. It consists of 12.791 class-balanced samples labelled with one of six labels according to human fact cheking of Politifact.com. We binarise the dataset by labelling all "pants-fire", "false" and "barely true" samples as FAKE, and the rest ("half-true", "mostly-true", and "true") as RELIABLE.

In order to later (§3) compare feature importances across samples, we split the LIAR trainset into two parts, LIAR$_1$ and LIAR$_2$, the latter serving as an in-domain dataset to compare with. We also make a randomly relabelled version of the LIAR$_2$ set where the class labels are randomly shuffled, denoted LIAR$_2$ *raznom*. We use this dataset as a sanity check to compare LIAR$_2$ results with.

FAKENEWSCORPUS[2] is a large dataset with scraped news articles of different types, where the class of an article is defined by its source. For the purpose of our study, we selected articles of type "fake" and "reliable"[3] for our two-class setting. We sampled the first 25.000 articles of each type. The samples were shuffled and split into a train and test set with a 33% test set size.

The KAGGLE dataset [4] is from a Kaggle competition set up by the UTK Machine Learning Club. We work only with the train set posted on the webpage since the labels for the test set were never published. The dataset consist of 6938 "reliable" and 6998 "unreliable" samples. We split it into a train and test set with a 33% test set size.

Common for all datasets, and for the following experiments, we encode FAKE samples as label 0, and RELIABLE as 1. We also only work with the news article/statement content and do not use any metadata such as author or title in our experiments.

2.2 Models

We explore a standard architecture for fake news detection based on long short-term memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997), as well as an easily interpretable linear model (logistic regression) based on n-gram features. The LSTM architecture is similar to those used in state-of-the-art fake news detectors (Hanselowski et al., 2018; Popat et al., 2018; Ruchansky et al., 2017), and the linear model makes it easier for us to analyse the datasets.

For the **LSTM models**, we train both a bi-LSTM and a simple LSTM. The hyperparameter settings of both networks is based on the bi-LSTM experiments by Wang (2017), such that results on the LIAR dataset are comparable. Like Wang, we also used 300-dimensional pre-trained word2vec embeddings from Google News. For the bi-LSTM we train with categorical cross-entropy loss and make prediction with softmax activa-

[2] https://github.com/several27/FakeNewsCorpus

[3] The class is based on the trustworthiness of the source and we disregard grey-zone labels, such as "satire".

[4] https://www.kaggle.com/c/fake-news/data

Figure 1: LSTMVis visualisation of decision-making in our LSTM fake news detector. The top row of labels shows the predictions by our fake news detector, while the bottom row shows the true labels (FAKE=0 and RELIABLE=1).

tion. The simple LSTM is trained with binary cross-entropy loss and make timedistributed predictions with sigmoid activation. The purpose of the simple LSTM is to be able to visualize and inspect learned patterns using the LSTMVis visualisation tool (Strobelt et al., 2018). The networks are trained on the LIAR dataset and tested in-domain as well as the two other datasets.

For the **linear models**, the samples are vectorized with a TF-IDF vectorizer including lowercased uni–five-grams and a maximum of 5.000 most frequent features. To find the most predictive features, we use a shared vocabulary across samples and rank the features by coefficients in the direction of the fake class.

3 Results

The results of the **LSTM models** are presented in Table 1. For comparison, Wang (2017) report in-domain six-class accuracy. Our in-domain accuracy six-class accuracy is 25.4, whereas Wang report 23.3. The main observation is that the performance of our fake news detectors drops from > 0.6 to below random. This, in other words, means that the model trained on LIAR knows *nothing* about what is fake, and what is not, on the other datasets. We take this as a strong indicator that our models have fitted patterns that are confounds, not indicators, of posts being fake.

To analyse this further, we compare the feature importance rankings of models trained on two different splits of the LIAR training set, as well as to models trained on KAGGLE and FAKENEWSCORPUS, as well as a random relabelling of the LIAR dataset. Since we cannot easily extract feature importance from our LSTM models,[5] we do so from our linear models trained on n-grams. The in-domain accuracies of the linear models are 55., 59., 95., and 34. for LIAR_1, LIAR_2, KAGGLE and FAKENEWSCORPUS, respectively. This reveals large differences in the ease of the task.

Table 2 show a quantitative evaluation of the similarity of the importance ranking of features induced by the different linear models. We use three standard metrics for comparing rankings: mean reciprocal rank (MRR), Kendall's τ, and Spearman's ρ. The results are remarkable: The correlation across samples, e.g., between the feature importances of the baseline model trained on LIAR and the feature importances of the model trained on KAGGLE and FAKENEWSCORPUS, are as low as the correlation between the baseline and the model trained on a randomly relabelled LIAR. Since the latter induces a random ordering of features, this means the rankings have nothing in common.

The low similarity of the induced rankings suggests that our models overfit to sample-specific correlates instead of learning patterns that generalise well. In order to verify this, we inspect the most important features across the four models in Table 3, excluding the model trained on a random relabelling of LIAR, whose ranking is random. It is clear from the most important features that many of them do not seem intuitively related to fakeness at all. The second-most important feature in the LIAR_1 model is the trigram *this is the*; the second-most important feature in the FAKENEWSCORPUS model is the bigram *next by*. Overall the rankings are very different, and none of the top-20 features overlap. When we *inspect* our LSTM models, we see that they also react to seemingly arbitrary input patterns. See the visualisation in Figure 1 obtained using LSTMVis[6]. This is a non-fake post, labelled fake. Note that the LSTM

[5] See Rei and Søgaard (2018) for work along these lines.

[6] http://lstm.seas.harvard.edu

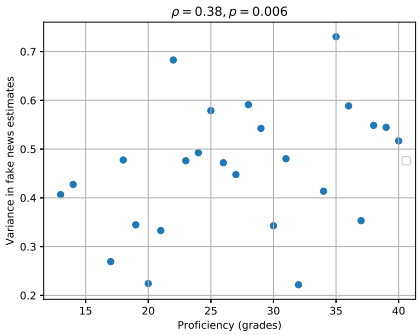

Figure 2: Correlation between grades and variance in predictions of our fake news detector

fake news detector finds evidence that the post is non-fake, in the pattern *there have . . . cities*, which seems non-intuitive.

4 Discussion

Dangerous Biases Pérez-Rosas et al. (2018) showed that readability features, indicating how skilled the writer of a post is, are the most predictive of whether a post is fake. We think this is a bug, not a feature. Our fake news detectors should not be more likely to predict less-skilled authors to write fake posts. Pérez-Rosas et al. (2018) explicitly used readability features in their models, but of course it does not follow that LSTM models exhibit similar biases.

In order to evaluate this, we ran our fake news detector trained on LIAR on a corpus produced by language learners, namely, the First Certificate in English (FCE) Corpus (Yannakoudakis et al., 2011). The FCE corpus comprises 1244 exam scripts from Cambridge ESOL First Certificate in English tests, including exam scores and native languages, as well as the error corrections by an examiner. We use the students' texts from the official test set (194 exam scripts by 97 unique authors, two scripts per author).

We are interested in the likelihood that our fake news detector accidentally labels a post written by an author with a certain level of English proficiency, as fake. We therefore correlate, per student, the variance in predicted confidence that her scripts are fake posts, with the student's proficiency level. We plot these in Figure 2. The correlation is highly significant (p <0.001) and quite strong, with Spearman's $\rho = 0.38$.

Is Knowledge-Free Fake News Detection Possible? In order to answer this, we need to distinguish between general world knowledge and specific knowledge about the topic at hand. With sufficient general world knowledge, we can say certain claims about e.g. politicians are *hard to believe* or *unlikely to have happened that way*; but without specific knowledge of what actually happened, we do not know with confidence what is fake or not.

Nevertheless, in both psychological literature (Pennebaker, 2011) and NLP tasks, such as opinion spam (Ott et al., 2011), linguistic markers for deception or satire (Rubin et al., 2016) are reported. If language reflects whether we lie, then we might expect this to spill over in fake news, but existing datasets are too small to leverage the subtlety of the signal.

Related Work The work most related to ours is O'Brien et al. (2018). They train a convolutional neural network for text classification on a Kaggle dataset and evaluate it on a held-out topic (from the same dataset), reporting good generalisation to this topic. They use back-propagation to say which features contribute most to the classification. In contrast, we evaluate a fake news detection model across different samples, with different distributions of authors, reporting *poor* generalization. This result shows that the confounds detected by such models do not, in general, correlate with whether stories are fake. O'Brien et al. (2018) claim their detector "captures subtle differences in the language of fake and real news." In contrast, we argue that detectors trained in this way mostly seem to capture spurious correlations. Finally, we analyse the bias in our detector's representations by correlating it with authors' language skills. Stylometric features were useful for discriminating hyperpartisan texts from mainstream, but were found insufficient for detecting fake news (Potthast et al., 2018), which is also why Potthast et al. support knowledge-based approaches for fake news detection.

5 Conclusion

We have shown that knowledge-free fake news classifiers do not generalise to other samples. We have performed thorough analyses to reveal that such models rely on sample-specific correlations and that models may even be unfortunately biased with respect to demographic variables.

References

Richard Ernest Bellman. 1957. *Dynamic Programming*. Princeton University Press.

Andreas Hanselowski, PVS Avinesh, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Nicole O'Brien, Sophia Latessa, Georgios Evangelopoulos, and Xavier Boix. 2018. The language of fake news: Opening the black-box of deep learning based detectors. In *NeurIPS*.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 309–319. Association for Computational Linguistics.

James Pennebaker. 2011. *The secret life of pronouns*. Bloomsbury Press.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240.

Marek Rei and Anders Søgaard. 2018. Zero-shot sequence labeling: Transferring knowledge from sentences to tokens. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 293–302.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806. ACM.

Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M. Rush. 2018. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 24:667–676.

William Yang Wang. 2017. liar, liar pants on fire: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 180–189. Association for Computational Linguistics.