





Learning Deep Fake-News Detectors from Scarcely-Labelled News Corpora

P. Zicari¹^a, M. Guarascio²^b, L. Pontieri²^c and G. Folino²^d

¹*DIMES, University of Calabria, Via P. Bucci, 87036 Rende (CS), Italy*

²*Institute of High Performance Computing and Networking (ICAR-CNR), Via P. Bucci, 87036 Rende (CS), Italy*

Keywords: Fake News Detection, Deep Learning, Pseudo-Labeling, Text Classification.

Abstract: Nowadays, news can be rapidly published and shared through several different channels (e.g., *Twitter*, *Facebook*, *Instagram*, etc.) and reach every person worldwide. However, this information is typically unverified and/or interpreted according to the point of view of the publisher. Consequently, malicious users can leverage these unofficial channels to share misleading or false news to manipulate the opinion of the readers and make fake news viral. In this scenario, early detection of this malicious information is challenging as it requires coping with several issues (e.g., scarcity of labelled data, unbalanced class distribution, and efficient handling of raw data). To address all these issues, in this work, we propose a Semi-Supervised Deep Learning based approach that allows for discovering accurate and effective Fake News Detection models. By embedding a BERT model in a pseudo-labelling procedure, the approach can yield reliable detection models also when a limited number of examples are available. Extensive experimentation on two benchmark datasets demonstrates the quality of the proposed solution.

1 INTRODUCTION

In recent times, (unofficial) social-media channels, such as *Twitter*, *Facebook*, and *Instagram*, have been exploited to widespread false information and influence the opinion of the people. This phenomenon took different forms over time: disinformation, click-bait, misinformation, and deceptive news are some examples, just to cite a few (Zhou and Zafarani, 2020).


The exacerbation of this problem has attracted the attention of researchers and practitioners, especially because of the suspicion that several important recent events (e.g., the 2016 US election, the Brexit referendum, and the Vax campaign for the COVID-19 pandemic emergency) were influenced by the diffusion of misleading information. As a matter of fact, massive amounts of possibly manipulated news are nowadays made available through traditional main media, online social systems, and personal broadcasting systems.


In this scenario, assessing the veracity and truth


of news is a crucial task that can benefit from recent advances in the field of Artificial Intelligence (AI) and Machine Learning (ML). As this task is time-consuming, expensive, and unfeasible on large data streams generated on the Web, AI-Based tools represent an effective solution to automate the identification of malicious information by reducing the intervention of trusted professionals and specialists.


In the literature, fake news detection was traditionally tackled as a problem of text classification (Liu et al., 2019), discriminating between real and fake news documents. However, training detection models to effectively recognize malicious information requires addressing many complex issues. First, a reliable solution should be able to handle low-level raw data frequently affected by noise, as the channels used to spread fake news typically allow for sharing only short text. In addition, the number of labelled training instances is limited; indeed, the labelling phase is a difficult and time-consuming task manually performed by domain experts. Finally, malicious contents represent only a limited portion of the data; then, the training set will exhibit an unbalanced distribution that makes it more difficult the learning phase of the model.

To overcome the limitations of traditional ap-

^a  <https://orcid.org/0000-0002-9119-9865>

^b  <https://orcid.org/0000-0001-7711-9833>

^c  <https://orcid.org/0000-0003-4513-0362>

^d  <https://orcid.org/0000-0002-8139-3445>

proaches, in this work, we define a semi-supervised deep learning-based approach able to discover effective fake news detection models when a limited number of instances are available for the training phase. The adoption of the Deep Learning (DL) paradigm looks like a natural solution to this problem, as DL techniques permit the learning of accurate classification models also from raw data without requiring heavy intervention by data-science experts (Guarascio et al., 2018). Basically, these DL models are structured according to a hierarchical architecture (consisting of several layers of base computational units, i.e., the artificial neurons are stacked one upon the other), allowing for learning features at different abstraction levels to represent raw data. In recent years, several sophisticated DL-based language models were proven excellent at learning (if trained against large document corpora) general hierarchical text representations (Liu et al., 2020), capturing the structure and semantics of the natural language. The language modelling abilities of such pre-trained models are often exploited in fine-tuning schemes to adapt their internal hierarchical representations of text data to specific text classification tasks.

In the fake news detection approach that is being proposed here, a pre-trained instance of BERT model (Devlin et al., 2019) is exploited as a backbone for classifying (as either fake or not) short news documents coming from a specific domain. However, instead of simply trying to fine-tune the pre-trained BERT model for this classification task, we propose embedding it into a self-training scheme. This allows us to fully exploit the unlabelled data available (along with their associated pseudo labels) to complement the training examples equipped with ground-truth class labels. Extensive experiments conducted on two different datasets confirmed the effectiveness of our approach in discovering accurate enough classification models even when the fraction of labelled data is relatively small. To the best of our knowledge, this work has been the first attempt in the field to combine the usage of a (unsupervisedly) pre-trained BERT model with a (pseudo-label based) self-training scheme.

The rest of this paper is organised as follows: Section 2 surveys some relevant works related to our research; Section 3 contains some background information, possibly useful for better understanding our proposed solution; in Section 4, an overview of the pseudo-labelling based scheme is provided; the experimentation results are illustrated in Section 5; finally, Section 6 concludes the work and proposes some future possible research directions.

2 RELATED WORKS

In the literature, three main types of fake news detection have been proposed: (i) knowledge-based, (ii) content-based and (iii) context-based.

Fake news detection based on knowledge is named *fact checking*, as it adopts the approach of checking the authenticity of news by comparing the information with documents or web resources extracted from the semantic web, linked open data and/or information retrieval. Content-based detection techniques analyse content and writing style to identify fake news and are based on Machine and Deep Learning methods. Finally, context-based detection approaches combine the news content with other information, e.g., the source, the author, the website, the topic, the propagation path and the speed of dissemination.

Content-based approaches to fake news detection constitute the prevalent kind of solutions in the field due to their broader applicability. Indeed, it is not easy to obtain high-quality integrated information from heterogeneous sources. Even though a large part of the content-based methods proposed so far rely on traditional supervised learning methods, it is important to remark that obtaining appropriate fake-news detection models via supervised learning entails gathering large amounts of reliable (labelled) data, which is time-consuming, expensive and requires specific topic knowledge. Thus, providing fake news detection systems with the ability to also exploit unlabelled data via semi-supervised learning mechanisms is necessary to suitably deal with real-life application scenarios where only small fractions of news documents are provided with a fake/normal class label.

In what follows, we survey some major semi-supervised approaches for the discovery of content-based classification models for fake news detection.

In (Rout et al., 2017), the authors compare four methods for detecting deceptive and fake opinion reviews: co-training, expectation maximisation, label propagation and spreading, and positive unlabelled learning. Co-training is a technique that exploits different views of the dataset, where each view is a distribution of features representing the data; the basic idea is to train two classifiers on each view and then classify instances on the unlabelled category to enlarge the training set. Expectation maximization consists of two steps: the learning of the algorithm with the conjunction of the labelled and predicted labelled sets (Expectation step) and the prediction of the labels of the unlabelled set (Maximization step). Label propagation and spreading use graph-based algorithms for learning: the graph is constructed by ordering suitable

vector features based on a suitable similarity metric, such as Manhattan distance or Euclidean distance, on both labelled and unlabelled nodes; label information is spread across the graph dynamically until all nodes are labelled. Positive unlabelled learning refers to a specific binary classification problem characterised by the constraint that only positive labelled data are available together with unlabelled data, and the classifier has to identify hidden positives from the set of unlabelled examples when negative training data is not supplied or available.

The work in (Guacho et al., 2020) proposes a semi-supervised fake detection classifier consisting of three phases: building the tensor-based embeddings representation of the article text; constructing a k-NN graph of proximal embeddings; and propagating the beliefs by using the *FaBP* (Fast Belief Propagation) algorithm. A similar approach, based on a graph-based semi-supervised fake news detection algorithm, is proposed in (Benamira et al., 2019), exploiting document embedding, graph inference for the representation of articles, and a graph neural network-based classifier.

In (Meel and Vishwakarma, 2021b), a semi-supervised temporal ensemble model is learned by using a Convolutional Neural Network (CNN) as a reference architecture for training the base models against the headline and the body of the news. The underlying idea of the temporal ensembling technique (Laine and Aila, 2016) is that different prediction outputs of all previous epochs can be aggregated in order to furnish a collaborative prediction which proved to be more accurate and thus better suitable for inferring pseudo labels. Indeed, the ensemble predictions of unknown labels accumulated in several training epochs perform better than the last epoch prediction.

In (Meel and Vishwakarma, 2021a), the same authors have also proposed a semi-supervised fake news detection technique based on GCN (Graph Convolutional Networks) trained with limited amounts of labelled data. The proposed solution consists of three stages: extracting an embedded representation from the news text by using *GloVe*, constructing a similarity graph using *Word Mover's Distance (WMD)*, and finally leveraging a Graph Convolution Network to address the binary classification task in a semi-supervised paradigm.

In (Dong et al., 2019), the authors introduce a novel deep two-path semi-supervised learning (DTSL) model composed of three convolutional subnets. The first is trained by using a supervised learning scheme, while the second is trained against unlabelled data in an unsupervised fashion. An addi-

tional shared CNN is used to propagate low-level features to the two former networks. The loss function is computed by weighting two components: a standard cross-entropy loss function to evaluate the loss for labelled inputs only and the mean squared error of the two output path predictions in order to penalize different predictions for the same training input.

To the best of our knowledge, our current work has been the first attempt to combine (language-model) pre-training and (pseudo-label based) self-training in order to train a powerful (BERT-based) deep model to discriminate fake news from genuine ones. As a matter of fact, the idea of resorting to an “old-fashion” pseudo-labelling approach was inspired by the results of the empirical analysis in (Cascente-Bonilla et al., 2021), where it was shown that such an approach is competitive with state-of-the-art semi-supervised DL methods (leveraging consistency regularization mechanisms) while being more resilient to out-of-distribution samples in the unlabeled set.

3 BACKGROUND

This section provides some background information on the specific neural network classifier used in the proposed approach and on the usage of pseudo-labels in a semi-supervised learning scenario.

The current implementation of our technique works on news texts gathered from different web sources by exploiting a pre-trained *BERT* (Bidirectional Encoder Representations from Transformers) model (Devlin et al., 2019) as a neural-network embedder. Essentially, *BERT* is a transformer-based neural architecture able to process natural language. It is trained through an algorithm including two main steps, named *Word Masking* and *Next Sentence Prediction (NSP)*, respectively. In the former step, a percentage of the words composing a sentence is masked, and the model is trained to predict the missing terms by considering the word context, i.e., the terms that precede and follow the masked one. Then, the model is fine-tuned by considering a further task to understand the sentences' relations. An overview of this learning procedure is depicted in figure 1. In our framework, we adopt a *BERT* instance pre-trained on Wikipedia pages, then improved by using an iterative self-training scheme (see Section 4) described below.

A Pseudo-Labeling approach is used in this work to map a number of unlabelled data instances, sampled from a given instance bucket, with pseudo labels assigned to them by a classification model, which is iteratively trained against a growing collection of both originally-labelled examples and pseudo-

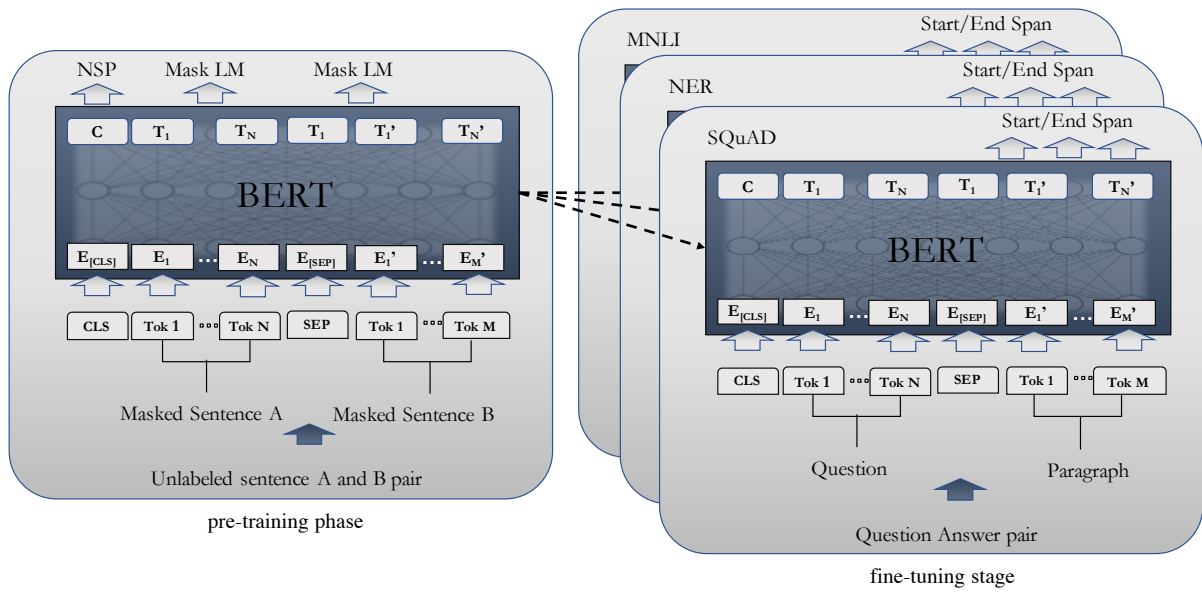


Figure 1: BERT Learning scheme.

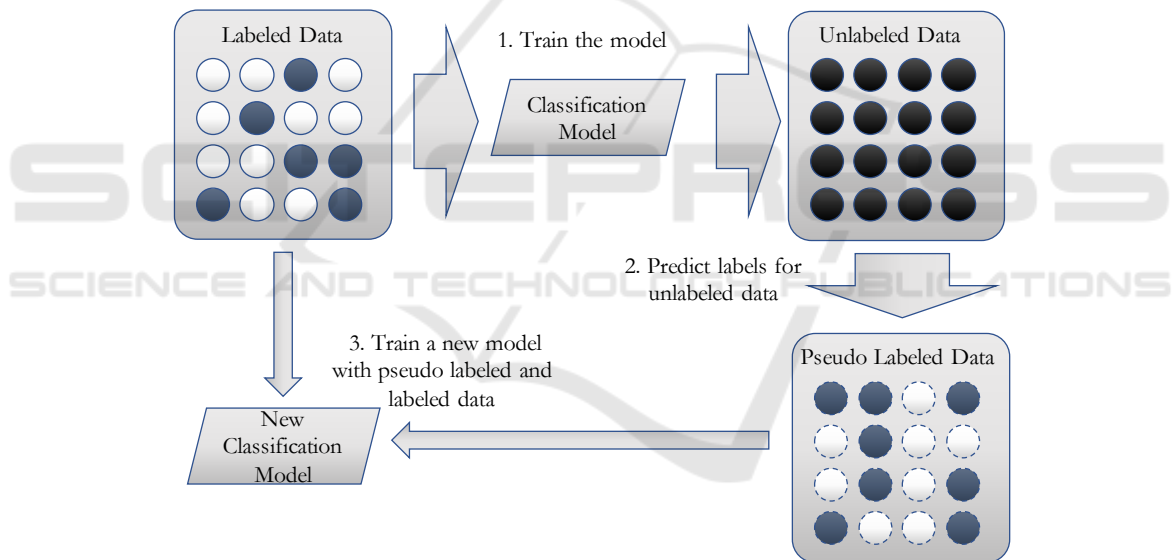


Figure 2: Pseudo-labelling approach to train a fake news classifier: high-level simplified view.

labelled ones. This process is repeated (in a self-training cycle) until some suitable stop criterion is satisfied (e.g., *the maximum number of iterations, loss convergence*, etc.). This approach is sketched in Figure 2, where the initial set of labelled data is enriched with unlabelled data while using the classifier trained on the labelled data to assign “artificial” labels to unlabelled ones. Then, a new version of the classifier is built by using both the originally-labelled data and the newly automatically-labelled ones.

Most of the pseudo-labelling techniques proposed in the literature are related to the image classification task. In particular, using pseudo-labels in the semi-

supervised learning of neural networks was originally proposed in (Lee, 2013), where un-labelled data are provided with pseudo-labels by just picking up the class which has the maximum predicted probability. By minimizing the conditional entropy of class probabilities for unlabeled data, the proposed method is demonstrated to be equivalent to the Entropy Regularization, thus favoring a low-density separation between classes.

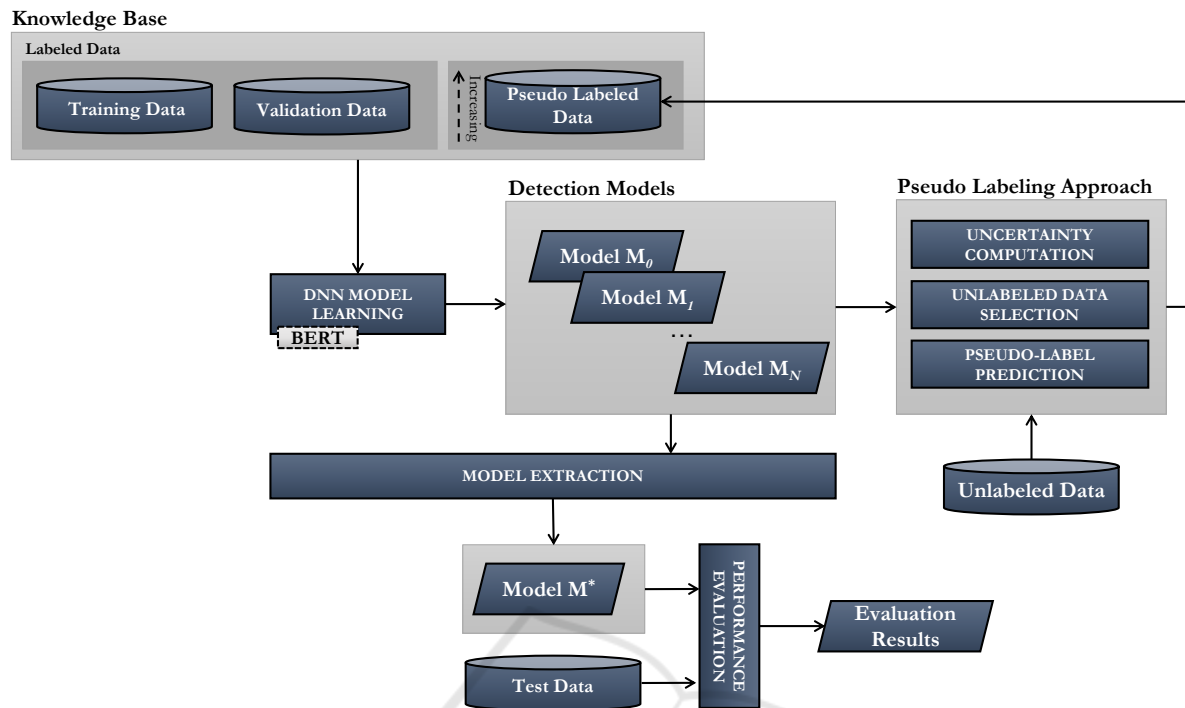


Figure 3: Conceptual architecture of the proposed semi-supervised DL framework for fake news detection.

4 FAKE NEWS DETECTION FRAMEWORK

This section illustrates the architecture of the framework developed to detect fake news and the details of the two alternative semi-supervised learning strategies that can exploit in order to cope with the scarcity of labelled training instances.

4.1 Conceptual Architecture of the Framework

Figure 3 illustrates a conceptual system architecture of the framework that we are proposing for the discovery, application and evaluation of deep fake news classifiers. This architecture was specifically devised to face the problem that only a small portion of the available examples of news data is associated with a class label so most of the examples are unlabelled.

As shown in the top-left of the figure, this problem is overcome by progressively enriching the given labelled data instances with novel pseudo-labelled tuples. At the very beginning of this iterative learning process, a preliminary classification model M_0 is built (by the *DNN model learning* module), by reusing a pre-trained instance of BERT as a backbone. This fine-tuning task is performed by only using the given

labelled news data instances, split as usual into a training set and a validation set.

Afterwards, an iterative process is followed, which consists of two phases. In the first phase, the generated model is exploited (by the *Pseudo Labeling Approach* module) to estimate the class of unlabelled data instances, and to assign an artificial class label to some of them eventually; the latter data instances are selected (by the *unlabelled data selection* module) according to one of the two strategies described in Subsection 4.2. In the second phase, the batch of (pseudo-) labelled data instances obtained in the former phase is added to the training set, and exploited, together with those already available before, to train a new version of the classification model (e.g., M_1 at iteration 1, M_2 at iteration 2 and so on), which is stored in the *Detection Model Repository*. These phases are iterated until no new element of the pseudo-labelling set meets the constraints defined by the selection strategy (e.g., until the probability of the model correctly predicting a tuple goes under a given threshold) or until there are no more available unlabelled instances.

Finally, the *Model Extraction* module selects one of the models generated during the different iterations, to be used as the final classifier for detecting incoming fake news. In the current implementation, the model obtained at the last self-training iteration is returned as a result, but different implementations

Algorithm 1: Pseudo-code for self-training the model with the Pseudo-Labeling algorithm.

```

Input      : Initial Training Set ( $ITrS$ );
              Validation Set ( $VS$ );
              Unlabelled Set ( $ULS$ ); // Unlabelled instances, for which pseudo labels are produced
Parameters:  $strategy \in \{str\_bestK, str\_thr\}$ ;
               $thr$ ; //maximal uncertainty threshold (to be used when setting  $strategy = str\_thr$ )
               $k$ ; //maximum number of instances (to be used when setting  $strategy = str\_bestK$ )
Output    : A news classification model  $M$ 
1  $Train(M, ITrS, VS)$  // train the classifier  $M$  using  $ITrS$  and  $VS$ 
2  $TrS = ITrS$  // Current Training Set
3  $PsS = \emptyset$  // Pseudo labelled Set
4  $newPseudoLabel = True$ 
5 while  $|ULS| > 0$  AND  $newPseudoLabel$  do
6    $newPseudoLabel = False$ 
7    $U = ComputeUncertaintyScores(M, ULS)$  //  $U$  is an ordered list of pairs of the form  $(x, u)$  such that
       $x \in ULS$  and  $u \in \mathbb{R}^+$  is a score quantifying how much  $M$  is uncertain in making a prediction for  $x$ 
8    $X = SelectUnlabelled(ULS, U, strategy, k, thr)$  // Select the unlabelled data to be pseudo labelled
9   if  $|X| > 0$  then
10     $newPseudoLabel = True$ 
11     $PsS = \{(x, M(x)) \mid x \in X\}$ ; // Generate a bunch of pseudo-labelled instances, by assigning a
      predicted label to each of the selected unlabeled instances in  $X$ 
12     $ULS = ULS \setminus X$ 
13     $TrS = TrS \cup PsS$ 
14     $Train(M, TrS, VS)$  // Train the model  $M$  from scratch using the tr. set  $TrS$  and the val. set  $VS$ 
15  end
16 end
17 return  $M$ 

```

could also be considered (that will be explored in future work). The *Performance Evaluation* module returns different evaluation metrics used in the experimental section.

A more detailed description of the proposed self-training strategy, based on pseudo-labelling, is provided in the next subsection, in the form of an algorithm (named Algorithm 1).

4.2 Pseudo-Labeling Algorithm and Unlabelled-Data Selection Strategies

The pseudo-code in Algorithm 1 explains in detail how the proposed training algorithm works.

Given as input an initial Training Set ($ITrS$), a Validation Set (VS) and a set of unlabelled instances (ULS) coming from a stream of news, an initial model M is trained using the labelled instances of the two sets $ITrS$ and VS . The learning process goes on through multiple training rounds, using both the manually-labelled data initially stored in $ITrS$ and VS , and the pseudo-labelled data added to PsS —i.e. data without true labels that are labelled based on the predictions returned by the model obtained at former iterations. More precisely, the following operations

are performed until no new pseudo-labelled instances are added to the current training set (TrS).

First, for each instance of the unlabelled set, an uncertainty score U is computed, which is meant to estimate how much the prediction returned by model M for x is uncertain. In principle, different uncertainty estimation methods (Mena et al., 2021) could be adopted for this aim. In the implementation of the framework that was employed in the experimental analysis of Section 5, the uncertainty scores are simply derived from the highest class membership probability returned by M for x (the closer this probability is to 0.5 the higher the uncertainty degree).

Then, a subset X of instances taken from ULS are selected for being pseudo labelled by preferring those ones on which the current model M seems to be less uncertain. Two different strategies can be adopted to make this selection step (described later on), and they can be controlled through the parameter *strategy* of Algorithm 1.

The selected instances in X are automatically assigned a pseudo label with the help of the current model M , and put into a new temporary (Pseudo-Label) set PsS . All of these pseudo-labelled instances are added to the current training set TrS , while re-

Table 1: Main features of the *PolitiFact* and *GossipCop* dataset.

Dataset	# Articles	# Classes	Vocabulary size	# Words per article			
				Avg.	Median	Q1	Q3
<i>PolitiFact</i>	814	2	60,870	817.5	199.5	66.5	530.7
<i>GossipCop</i>	4,719	2	149,557	349.0	205.0	109.5	323.0

moving all the instances of X from the set ULS of unlabelled data instances.

Finally, a new model M is trained from scratch over the (augmented) training set TrS , still using VS as validation set, after initializing the weights of all the layers of M but the last (which is initialised randomly) with the weights of the pre-trained BERT model. It is worth noting that, in order to curb the risk of confirmation bias and of concept drifts, we do not adopt a sort of incremental training scheme where M is initialised with a copy of the model obtained at the previous iteration of the self-training loop (Steps 6-18 of Algorithm 1). Indeed, restarting model parameters before each self-training cycle was identified in (Cascante-Bonilla et al., 2021) as a key to the success of pseudo-labelling approaches to the discovery of deep models.

At the end of the self-training loop, the current classification model M is returned, which is the one obtained at the last self-training iteration. It is worth noting that more sophisticated model extraction schemes could be devised that take advantage of classification models obtained at other self-training iterations (e.g., possibly exploiting ensemble learning to combine multiple models), but this is left to future work.

Selection Strategy (Parameter Strategy of Algorithm 1). Two alternative strategies can be adopted (by the *Pseudo-label Strategy Manager* module) for selecting which unlabelled data are promoted to pseudo-labelled ones, in order to obtain an improved version of the fake news classifier.

One strategy (chosen when setting $strategy = str_thr$ in Algorithm 1), simply consists in comparing the uncertainty score of an unlabelled data instance to a given maximal threshold thr . The subset of samples in the unlabelled set (ULS) to be included in the Training set (TrS) is built by selecting the instances for which the lastly trained model M returns a prediction with an uncertainty score lower than thr .

The second strategy (chosen when setting $strategy = str_bestK$ in Algorithm 1) consists in ranking of the instances in ULS based on their associated prediction-uncertainty scores and eventually selecting the k ones of them achieving the lowest scores.

In both cases, each instance x , among those se-

lected as described so far, is artificially assigned a (pseudo-)label that refers to the class for which the model returned the highest class membership probability on x .

5 EXPERIMENTAL RESULTS

5.1 Datasets and Parameters

This subsection describes the parameters used in our framework and the datasets used to assess its effectiveness in detecting fake news.

The learning model employed in the performed tests is based on a *BERT* layer followed by a *Dropout* layer for regularisation and a final dense layer with a sigmoid activation layer. The *BERT* implementation presents a vector of hidden size of 768, and 12 attention heads. The used model is pre-trained for the English language on *Wikipedia* and *BooksCorpus*, after a normalisation phase.

The following parameters were used in BERT: Number of Epoch = 30; Batch size = 32; Learning Rate = $3e - 5$; Dropout = 0.1; the Binary Cross entropy as loss function and the chosen optimiser was AdamW, a stochastic optimisation method that modifies the typical implementation of weight decay in Adam, by decoupling weight decay from the gradient update.

It is worth recalling that two strategies can be employed in our framework for iteratively selecting unlabelled data to be pseudo-labelled, described in Section 4: strategy str_thr (which relies on filtering candidates through a maximal uncertainty threshold) and strategy str_bestK (which extracts the “bottom- k ” tuples with the lowest prediction uncertainty).

A grid search was performed to choose the probability prediction thr in the case of the str_thr strategy and the number k of the best- k unlabelled data to be pseudo-labelled at each self training iteration for the str_bestK algorithm. Respectively, the values of $thr = 0.4$ and $k = 100$, and the values of $thr = 0.3$ and $k = 200$ were chosen for the *PolitiFact* and for the *GossipCop* dataset.

All the experiments of the next subsection were averaged over 30 runs. The validation set is used in

Table 2: Comparison of the pseudo-labelling strategies for the *PolitiFact* dataset: Accuracy, AUC, AUC-PR and F-measure for different percentages of the training set (5%, 10%, 15%, 20% and 25%).

Metric	Algorithm	5%	10%	15%	20%	25%
Accuracy	Baseline	0.68 ± 0.08	0.72 ± 0.05	0.76 ± 0.05	0.79 ± 0.05	0.80 ± 0.02
	<i>pseudo_k</i>	0.73 ± 0.07	0.78 ± 0.05	0.82 ± 0.05	0.86 ± 0.05	0.84 ± 0.04
	<i>pseudo_thr</i>	0.77 ± 0.02	0.83 ± 0.03	0.85 ± 0.03	0.86 ± 0.02	0.86 ± 0.04
AUC	Baseline	0.68 ± 0.08	0.71 ± 0.05	0.75 ± 0.05	0.79 ± 0.05	0.80 ± 0.03
	<i>pseudo_k</i>	0.72 ± 0.08	0.78 ± 0.04	0.82 ± 0.05	0.86 ± 0.05	0.84 ± 0.04
	<i>pseudo_thr</i>	0.77 ± 0.03	0.83 ± 0.03	0.85 ± 0.03	0.86 ± 0.02	0.86 ± 0.04
AUC_PR	Baseline	0.80 ± 0.05	0.82 ± 0.03	0.84 ± 0.03	0.86 ± 0.03	0.87 ± 0.02
	<i>pseudo_k</i>	0.83 ± 0.00	0.85 ± 0.00	0.88 ± 0.00	0.90 ± 0.00	0.89 ± 0.00
	<i>pseudo_thr</i>	0.85 ± 0.01	0.89 ± 0.02	0.90 ± 0.02	0.91 ± 0.01	0.90 ± 0.03
F1	Baseline	0.72 ± 0.07	0.76 ± 0.04	0.79 ± 0.04	0.81 ± 0.04	0.82 ± 0.02
	<i>pseudo_k</i>	0.76 ± 0.04	0.80 ± 0.05	0.83 ± 0.04	0.87 ± 0.05	0.85 ± 0.04
	<i>pseudo_thr</i>	0.79 ± 0.03	0.85 ± 0.04	0.86 ± 0.03	0.87 ± 0.03	0.87 ± 0.04

Table 3: Comparison of the pseudo-labelling strategies for the *GossipCop* dataset: Accuracy, AUC, AUC-PR and F-measure for different percentages of the training set (5%, 10%, 15%, 20% and 25%).

Metric	Algorithm	5%	10%	15%	20%	25%
Accuracy	Baseline	0.66 ± 0.03	0.69 ± 0.02	0.71 ± 0.03	0.70 ± 0.03	0.72 ± 0.02
	<i>pseudo_k</i>	0.69 ± 0.05	0.74 ± 0.02	0.74 ± 0.03	0.75 ± 0.03	0.75 ± 0.02
	<i>pseudo_thr</i>	0.72 ± 0.00	0.75 ± 0.00	0.75 ± 0.00	0.76 ± 0.00	0.76 ± 0.00
AUC	Baseline	0.66 ± 0.03	0.69 ± 0.03	0.70 ± 0.03	0.70 ± 0.03	0.72 ± 0.02
	<i>pseudo_k</i>	0.69 ± 0.05	0.73 ± 0.02	0.73 ± 0.03	0.74 ± 0.03	0.75 ± 0.02
	<i>pseudo_thr</i>	0.72 ± 0.01	0.74 ± 0.01	0.75 ± 0.03	0.76 ± 0.02	0.76 ± 0.02
AUC_PR	Baseline	0.78 ± 0.03	0.80 ± 0.02	0.81 ± 0.02	0.81 ± 0.02	0.82 ± 0.01
	<i>pseudo_k</i>	0.80 ± 0.03	0.83 ± 0.01	0.83 ± 0.02	0.84 ± 0.02	0.84 ± 0.01
	<i>pseudo_thr</i>	0.82 ± 0.00	0.84 ± 0.01	0.84 ± 0.02	0.84 ± 0.01	0.85 ± 0.01
F1	Baseline	0.69 ± 0.06	0.72 ± 0.01	0.73 ± 0.03	0.72 ± 0.05	0.75 ± 0.03
	<i>pseudo_k</i>	0.71 ± 0.08	0.77 ± 0.03	0.77 ± 0.03	0.77 ± 0.03	0.78 ± 0.02
	<i>pseudo_thr</i>	0.74 ± 0.04	0.78 ± 0.01	0.78 ± 0.03	0.78 ± 0.01	0.78 ± 0.01

the training process for selecting the best model during the different epochs.

The two datasets used for the experiments come from the *FakeNewsNet* data repository (Shu et al., 2018) (Shu et al., 2017). They respectively concern political and gossip news obtained by two fact-checking websites: *PolitiFact*¹ and *GossipCop*².

Table 1 reports the main characteristic of the two datasets: the overall number of articles, the vocabulary size and some statistics on the number of words per article (i.e., average, median, first and third quartile).

The performance of our methods and of the baseline is evaluated against the test set through four different metrics: the largely used Accuracy metric and some measures more appropriate for evaluating unbalanced datasets, i.e., AUC (Area Under the Curve), AUC-PR (Precision-Recall) and F-Measure.

¹<https://www.politifact.com/>

²<https://www.gossipcop.com/>

5.2 Experimental Validation of the Two Pseudo-Labeling Strategies

In this subsection, we evaluated our two strategies in comparison with the baseline when different percentages of labelled data (training set and validation set) are considered (5%, 10%, 15%, 20% and 25%), in order to consider the situation in which a few (costly) labelled data are available.

Tables 2 and 3 report the results of the comparison among the baseline (the traditional method consisting in fine-tuning the same pre-trained BERT model in a fully-supervised against the sole labelled data) and the two variants of the proposed approach (corresponding to the two different selection strategies *str_bestK* and *str_thr*) in terms of Accuracy, AUC, AUC-PR and F-measure for the *PolitiFact* and the *GossipCop* datasets, respectively.

Results highlight that the two proposed strategies obtain a substantial increment for all the metrics con-

Table 4: Delta increment of the pseudo-labelling strategies in comparison with the baseline for the *PolitiFact* dataset: Accuracy, AUC, AUC-PR and F-measure for different percentages of the training set (5%, 10%, 15%, 20% and 25%).

Metric	Algorithm	5%	10%	15%	20%	25%
Accuracy	<i>pseudo_k</i>	6.11%	7.85%	8.15%	8.46%	4.95%
	<i>pseudo_thr</i>	12.63%	15.40%	12.76%	9.09%	7.15%
AUC	<i>pseudo_k</i>	6.00%	8.79%	9.60%	8.94%	5.46%
	<i>pseudo_thr</i>	12.99%	16.46%	14.16%	9.57%	7.54%
AUC-PR	<i>pseudo_k</i>	4.16%	4.31%	4.95%	4.98%	3.37%
	<i>pseudo_thr</i>	7.02%	8.61%	7.51%	5.28%	4.42%
F1	<i>pseudo_k</i>	5.93%	4.82%	4.74%	6.56%	3.53%
	<i>pseudo_thr</i>	10.01%	11.27%	8.88%	7.20%	5.62%

Table 5: Delta increment of the pseudo-labelling strategies in comparison with the baseline for the pseudo-labelling strategies for the *GossipCop* dataset for different percentages of the training set (5%, 10%, 15%, 20% and 25%).

Metric	Algorithm	5%	10%	15%	20%	25%
Accuracy	<i>pseudo_k</i>	3.77%	6.51%	4.38%	6.47%	4.73%
	<i>pseudo_thr</i>	9.05%	8.16%	6.53%	7.89%	6.04%
AUC	<i>pseudo_k</i>	3.69%	6.12%	4.26%	6.19%	5.00%
	<i>pseudo_thr</i>	9.32%	7.49%	6.76%	7.76%	6.37%
AUC-PR	<i>pseudo_k</i>	2.34%	3.45%	2.53%	3.67%	2.82%
	<i>pseudo_thr</i>	5.35%	4.23%	3.81%	4.58%	3.59%
F1	<i>pseudo_k</i>	2.94%	6.42%	4.28%	6.64%	3.97%
	<i>pseudo_thr</i>	7.89%	8.51%	5.61%	7.74%	4.88%

sidered and for both datasets. The improvements are more evident for the *PolitiFact* dataset, which is characterised by a smaller number of samples (814), probably because the proposed self-training strategy is more efficient when the labelled data are very scarce.

Comparing the two employed strategies, the threshold-based method performs better than the other one for all the measures and methods, and the differences are more evident when labelled data are scarce (lower percentages).

Tables 4 and 5 show the performance improvements in terms of percentage increment for the different metrics of the proposed self-trained model strategies with respect to the baseline, when the *PolitiFact* and the *GossipCop* datasets are tested, respectively.

The increment performance results in tables 4 and 5 allow for a better understanding of the behaviour of the different proposed strategies when the percentages of labelled available data vary.

By analyzing the performance trend, it is possible to notice that, as expected, when the available labelled data increases, the two proposed pseudo-labelling strategies improve for all the performance metrics. However, when reaching the percentage value of 20% for the *PolitiFact* dataset and about 15% for the *Gossip* dataset, the value of the increment decreases. The reason behind this behaviour, probably, is due to the fact that when more labelled data are available, the model can be well trained by us-

ing only labelled data, while pseudo labels could deteriorate the performance by introducing incorrect labels. Moreover, it is evident that the performance of the pseudo-labelling strategies is also related to the specifically considered dataset.

6 CONCLUSIONS

In this work, we devised a semi-supervised deep learning based framework able to effectively detect fake news by coping with a number of relevant issues, i.e., noisy data, data scarcity, and unbalanced class distributions. The framework relies on building up a deep classifier via a novel combination of unsupervised (language-model) pre-training and self-training. Specifically, a BERT model pre-trained on Wikipedia data is embedded in an iterative self-training scheme where pseudo-labelled data are generated incrementally and exploited for fine-tuning the model.

Experiments conducted on two public datasets confirmed the quality of the approach in generating accurate models, also when a limited number of training examples are available in the early stages of the proposed semi-supervised method.

In future work, we plan to devise a novel ensemble strategy able to combine the different models trained over the pseudo-labelling iterations in order to im-

prove the accuracy of the fake news detector.

Moreover, we plan to evaluate the combination of our framework with more sophisticated uncertainty estimation methods, as well as to devise mechanisms for differentiating true labelled data from pseudo-labelled data in the self-training process, in order to reduce the risk of confirmation bias that may arise from computing traditional loss functions over pseudo labels.

ACKNOWLEDGEMENTS

This work was partly supported by the European Commission funded project "HumanE-AI-Net" (grant no. 952026) and by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU. Their support is gratefully acknowledged.

REFERENCES

- Benamira, A., Devillers, B., Lesot, E., Ray, A. K., Saadi, M., and Malliaros, F. D. (2019). Semi-supervised learning and graph neural networks for fake news detection. In Spezzano, F., Chen, W., and Xiao, X., editors, *ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining, Vancouver, British Columbia, Canada, 27-30 August, 2019*, pages 568–569. ACM.
- Cascante-Bonilla, P., Tan, F., Qi, Y., and Ordóñez, V. (2021). Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6912–6920.
- Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.
- Dong, X., Victor, U., Chowdhury, S., and Qian, L. (2019). Deep two-path semi-supervised learning for fake news detection. *CoRR*, abs/1906.05659.
- Guacho, G. B., Abdali, S., Shah, N., and Papalexakis, E. E. (2020). Semi-supervised content-based detection of misinformation via tensor embeddings. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '18*, page 322–325. IEEE Press.
- Guarascio, M., Manco, G., and Ritacco, E. (2018). Deep learning. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1-3:634–647.
- Laine, S. and Aila, T. (2016). Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.
- Lee, D.-H. (2013). Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*.
- Liu, C., Wu, X., Yu, M., Li, G., Jiang, J., Huang, W., and Lu, X. (2019). A two-stage model based on bert for short fake news detection. In Douligieris, C., Karagiannis, D., and Apostolou, D., editors, *Knowledge Science, Engineering and Management*, pages 172–183, Cham. Springer International Publishing.
- Liu, Z., Lin, Y., and Sun, M. (2020). *Representation learning for natural language processing*. Springer Nature.
- Meel, P. and Vishwakarma, D. K. (2021a). Fake news detection using semi-supervised graph convolutional network. *CoRR*, abs/2109.13476.
- Meel, P. and Vishwakarma, D. K. (2021b). A temporal ensembling based semi-supervised convnet for the detection of fake news articles. *Expert Systems with Applications*, 177:115002.
- Mena, J., Pujol, O., and Vitrià, J. (2021). A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. *ACM Comput. Surv.*, 54(9).
- Rout, J. K., Dalmia, A., Choo, K.-K. R., Bakshi, S., and Jena, S. K. (2017). Revisiting semi-supervised learning for online deceptive review detection. *IEEE Access*, 5:1319–1327.
- Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2018). Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.
- Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.
- Zhou, X. and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).