Early Detection of Fake News with Multi-source Weak Social Supervision

Kai Shu[1], Guoqing Zheng[2], Yichuan Li[3], Subhabrata Mukherjee[2], Ahmed Hassan Awadallah[2], Scott Ruston[3], and Huan Liu[3]

[1] Department of Computer Science, Illinois Institute of Technology, `kshu@iit.edu`
[2] Microsoft Research AI, {`zheng, submukhe, hassanam`}`@microsoft.com`
[3] Arizona State University, {`yichuan1, Scott.Ruston, huanliu`}`@asu.edu`

Abstract. Social media has greatly enabled people to participate in online activities at an unprecedented rate. However, this unrestricted access also exacerbates the spread of misinformation and fake news which cause confusion and chaos if not detected in a timely manner. Given the rapidly evolving nature of news events and the limited amount of annotated data, state-of-the-art systems on fake news detection face challenges for early detection. In this work, we exploit multiple weak signals from different sources from user engagements with contents (referred to as weak social supervision), and their complementary utilities to detect fake news. We jointly leverage limited amount of clean data along with weak signals from social engagements to train a fake news detector in a meta-learning framework which estimates the quality of different weak instances. Experiments on real-world datasets demonstrate that the proposed framework outperforms state-of-the-art baselines for early detection of fake news without using any user engagements at prediction time.

Keywords: Fake news · Weak social supervision · Meta learning

1 Introduction

Motivation. Social media platforms provide convenient means for users to create and share diverse information. Due to its massive availability and convenient access, more people seek and receive news information online. For instance, around 68% of U.S. adults consumed news from social media in 2018, a massive increase from corresponding 49% consumption in 2012[4] according to a survey by the Pew Research Center. However, social media also proliferates a plethora of misinformation and fake news, i.e., news stories with intentionally false information [27]. Research has shown that fake news spreads farther, faster, deeper, and more widely than true news [32]. For example, during the 2016 U.S. election, the top twenty frequently-discussed false election stories generated 8.7 million shares, reactions, and comments on Facebook, more than the total of 7.4 million shares of top twenty most-discussed true stories[5]. Widespread fake news can

[4] https://bit.ly/39zPnMd
[5] https://bit.ly/39xmXT7

Fig. 1: An illustration of a piece of fake news and related user engagements, which can be used for extracting weak social supervision. Users have different credibility, perceived bias, and express diverse sentiment to the news.

erode the public trust in government and professional journalism and lead to adverse real-life events. Thus, a timely detection of fake news on social media is critical to cultivate a healthy news ecosystem.

Challenges. First, fake news is diverse in terms of topics, content, publishing methods and media platforms, and sophisticated linguistic styles geared to emulate true news. Consequently, training machine learning models on such sophisticated content requires *large-scale annotated fake news data* that is egregiously difficult to obtain. Second, it is important to detect fake news early. Most of the research on fake news detection rely on signals that require a long time to aggregate, making them unsuitable for *early detection*. Third, the evolving nature of fake news makes it essential to analyze it with signals from multiple sources to better understand the context. A system solely relying on social networks and user engagements can be easily influenced by biased user feedback, whereas relying only on the content misses the rich auxiliary information from the available sources. In this work, we adopt an approach designed to address the above challenges for early detection of fake news with limited annotated data by leveraging weak supervision from multiple sources involving users and their social engagements – referred to as *weak social supervision*.

Existing work. Prior works on detecting fake news [17,33] rely on large amounts of labeled instances to train supervised models. Such large labeled training data is difficult to obtain in the early phase of fake news detection. To overcome this challenge, learning with weak supervision presents a viable solution [34]. Weak signals are used as constraints to regularize prediction models [29], or as loss correction mechanisms [6]. Often only a *single* source of weak labels is used.

Existing research has focused either on the textual content relying solely on the linguistic styles in the form of sentiment, bias and psycho-linguistic features [15] or on tracing user engagements on how fake news propagate through the network [33]. In this work, we utilize *weak social supervision* to address the above shortcomings. Consider the example in Figure 3. Though it is difficult

to identify the veracity considering the news content in isolation, the surrounding context from other users' posts and comments provide clues, in the form of opinions, stances, and sentiment, useful to detect fake news. For example, in Figure 3, the phrase "`kinda agree...`" indicates a positive sentiment to the news, whereas the phrase "`I just do not believe it...`" expresses a negative sentiment. Prior work has shown conflicting sentiments among propagators to indicate a higher probability of fake news [8,27]. Also, users have different credibility degrees in social media and less-credible ones are more likely to share fake news [28]. Although we do not have this information a priori, we can consider *agreement* between users as a weak proxy for their credibility. All of the aforementioned signals from different sources like content and social engagements can be leveraged as weak supervision signals to train machine learning models.
Contributions. We leverage weak social supervision to detect fake news from limited annotated data. In particular, our model leverages a small amount of manually-annotated clean data and a large amount of weakly annotated data by proxy signals from multiple sources for joint training in a meta-learning framework. Since not all weak instances are equally informative, the model learns to estimate their respective contributions for the end task. To this end, we develop a Label Weighting Network (LWN) to model the weight of these weak labels that regulate the learning process of the fake news classifier. The LWN serves as a meta-model to produce weights for the weak labels and can be trained by back-propagating the validation loss of a trained classifier on a separate set of clean data. The framework is uniquely suitable for early fake news detection, because it (1) leverages rich weak social supervision to boost model learning in a meta-learning fashion; and (2) only requires the news content during the prediction stage without relying on the social context as features for early prediction. Our contributions can be summarized as:

- **Problem.** We study a novel problem of exploiting weak social supervision for early fake news detection;
- **Method.** We provide a principled solution, dubbed MWSS to learn from Multiple-sources of Weak Social Supervision (MWSS) from multi-faceted social media data. Our framework is powered by meta learning with a Label Weighting Network (LWN) to capture the relative contribution of different weak social supervision signals for training;
- **Features.** We describe how to generate weak labels from social engagements of users that can be used to train our meta learning framework for early fake news detection along with quantitative quality assessment.
- **Experiments.** We conduct extensive experiments to demonstrate the effectiveness of the proposed framework for early fake news detection over competitive baselines.

2 Modeling Multi-source Weak Social Supervision

User engagements over news articles, including posting about, commenting on or recommending the news, bear implicit judgments of the users about the news and could serve as weak sources of labels for fake news detection. For instance, prior research has shown that contrasting sentiment of users on a piece of news article, and similarly different levels of credibility or bias, can be indicators of the underlying news being fake. However, these signals are noisy and need to be appropriately weighted for training supervised models. Due to the noisy nature of such social media engagements, we term these signals as *weak social supervision*.

To give a brief overview for the modeling, we define heuristic labeling functions (refer to Section 3) on user engagements to harvest such signals to weakly label a large amount of data. The weakly labeled data is combined with limited amount of manually annotated examples to build a fake news detection system that is better than training on either subset of the data. We emphasize that multiple weak labels can be generated for a single news article based on different labeling functions and we aim to jointly utilize both the clean examples as well as multiple sources of weak social supervision in this paper.

In this section, we first formulate the problem statement, and then focus on developing algorithms for the joint optimization of manually annotated clean and multi-source weakly labeled instances in a unified framework.

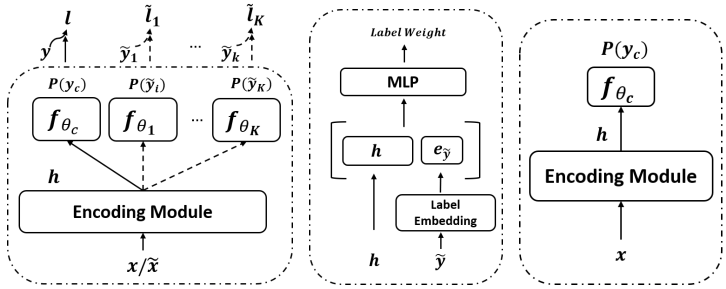
2.1 Problem Statement

Let $\mathcal{D} = \{x_i, y_i\}_{i=1}^{n}$ denote a set of n news articles with manually annotated clean labels, with $\mathcal{X} = \{x_i\}_{i=1}^{n}$ denoting the news pieces and $\mathcal{Y} = \{y_i\}_{i=1}^{n} \subset \{0,1\}^n$ the corresponding clean labels of whether the news is fake or not. In addition, there is a large set of unlabeled examples. Usually the size of the clean labeled set n is smaller than the unlabeled set due to labeling costs. For the widely available unlabeled samples, we can generate weak labels by using different labeling functions based on *social engagements*. For a specific labeling function $g^{(k)} : \mathcal{X}^{(k)} \to \tilde{\mathcal{Y}}^{(k)}$, where $\mathcal{X}^{(k)} = \{x_j^{(k)}\}_{j=1}^{N}$ denotes the set of N unlabeled messages to which the labeling function $g^{(k)}$ is applied and $\tilde{\mathcal{Y}}^{(k)} = \{\tilde{y}_j^{(k)}\}_{j=1}^{N}$ as the resulting set of weak labels. This weakly labeled data is denoted by $\tilde{\mathcal{D}}^{(k)} = \{x_j^{(k)}, \tilde{y}_j^{(k)}\}_{j=1}^{N}$ and often $n \ll N$. We formally define our problem as:

Problem Statement: Given a limited amount of manually annotated news data $\mathcal{D}$ and K sets of weakly labeled data $\{\tilde{\mathcal{D}}^{(k)}\}_{k=1}^{K}$ derived from K different weak labeling functions based on weak social signals, learn a fake news classifier $f : \mathcal{X} \to \mathcal{Y}$ which generalizes well onto unseen news pieces.

2.2 Meta Label Weighting with Multi-source Weak Social Supervision

Learning from multiple sources has shown promising performance in various domains such as truth discovery [4], object detection [13], etc. In this work, we have $K + 1$ distinct sources of supervision: clean labels coming from manual annotation and multiple sources of weak labels obtained from K heuristic labeling functions based on users' social engagements.



(a) MWSS Classifier (b) MWSS LWN (c) MWSS inference

Fig. 2: The proposed framework MWSS for learning with multiple weak supervision from social media data. (a) Classifier: Jointly modeling clean labels and weak labels from multiple sources; (b) LWN: Learning the label weight based on the concatenation of instance representation and weak label embedding vector. (c) During inference, MWSS uses the learned encoding module and classification MLP f_{w_c} to predict labels for (unseen) instances in the test data.

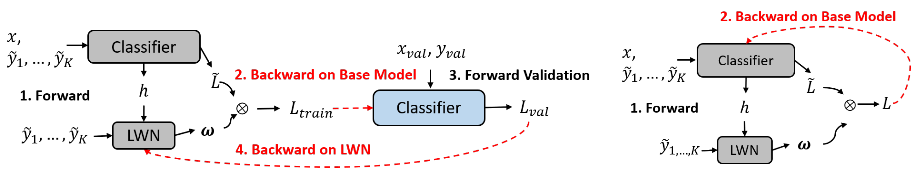
Our objective is to build an effective framework that leverages weak social supervision signals from multiple sources in addition to limited amount of clean data. However, signals from different weak sources are intrinsically noisy, biased in different ways, and thus of varying degree of qualities. Simply treating all sources of weak supervision as equally important and merging them to construct a single large set of weakly supervised instances tend to result in sub-optimal performance (as used as a baseline in our experiments). However, it is challenging to determine the contribution of different sources of weak social supervision. To facilitate a principled solution of weighting weak instances, we leverage meta-learning. In this, we propose to treat label weighting as a meta-procedure, i.e., building a *label weighting network* (LWN) which takes an instance (e.g., news piece) and its weak label (obtained from social supervision) as input, and outputs a scalar value as the importance weight for the pair. The weight determines the contribution of the weak instance in training the desired fake news classifier in our context. The LWN can be learned by back-propagating the loss of the trained classifier on a separate clean set of instances. To allow information sharing among different weak sources, for the fake news classifier, we use a shared feature extractor to learn a common representation and use separate functions (specifically, MLPs) to map the features to different weak label sources.

Specifically, let $h_{\theta_E}(x)$ be an encoder that generates the content representation of an instance x with parameters θ_E. Note that this encoder is shared by instances from both the clean and multiple weakly labeled sources. Let $f_{\theta_c}(h(x))$ and $\{f_{\theta_k}(h(x))\}_{k=1,..,K}$ be the $K+1$ labeling functions that map the contextual representation of the instances to their labels on the clean and K sets of weakly supervised data, respectively. In contrast to the encoder with shared parameters θ_E, the parameters θ_c and $\{\theta_k\}_{k=1,...,K}$ are different for the clean and weak

while *not converged* **do**
  1. Update LWN parameters α by descending $\nabla_\alpha \mathcal{L}_{val}(\theta - \eta \nabla_\theta \mathcal{L}_{train}(\alpha, \theta))$
  2. Update classifier parameters θ by descending $\nabla_\theta \mathcal{L}_{train}(\alpha, \theta)$

end

Algorithm 1: Training process of MWSS



(a) LWN Update (b) Classifier Update

Fig. 3: The illustration of the MWSS in two phases: (a) we compute the validation loss based on the validation dataset and retain the computation graph for LWN backward propagation; (b) the classifier update its parameters through backward propagation on clean and weakly labeled data. Note that h is the set of hidden representations of the input instances, ω is the weight for each pair of instances and labels and $\otimes$ is point-wise multiplication. Gray indicates the parameters from the last iteration, blue indicates temporary updates.

sources (learned by separate source-specific MLPs) to capture different mappings from the contextual representations to the labels from each source.

For training, we want to jointly optimize the loss functions defined over the (i) clean data and (ii) instances from the weak sources weighted by their respective utilities. The weight of the weak label $\tilde{y}$ for an instance x (encoded as $h(x)$) is determined by a separate Label Weighting Network (LWN) formulated as $\omega_\alpha(h(x), \tilde{y})$ with parameters α. Thus, for a given $\omega_\alpha(h(x), \tilde{y})$, the objective for training the predictive model with multiple sources of supervision jointly is:

$$\min_{\theta_E, \theta_c, \theta_1, \ldots, \theta_k} \mathbb{E}_{(x,y)\in\mathcal{D}} \ell(y, f_{\theta_c}(h_{\theta_E}(x))) + \sum_{k=1}^{K} \mathbb{E}_{(x,\tilde{y})\in\tilde{\mathcal{D}}^{(k)}} \omega_\alpha(h_{\theta_E}(x), \tilde{y}) \ell(\tilde{y}, f_{\theta_k}(h_{\theta_E}(x))) \tag{1}$$

where ℓ denotes the loss function to minimize the prediction error of the model. The first component in the above equation optimizes the loss over the clean data, whereas the second component optimizes for the weighted loss (given by $w_\alpha(.)$) of the weak instances from K sources. Figure 2 shows the formulation for both the classifier and LWN.

The final objective is to optimize LWN $\omega_\alpha(h(x), \tilde{y})$ such that when using such a weighting scheme to train the main classifier as specified by Eq. (1), the trained classifier could perform well on a separate set of clean examples. Formally, the following bi-level optimization problem describe the above intuition as:

$$\min_\alpha \mathcal{L}_{val}(\theta^*(\alpha)) \quad \text{s.t.} \quad \theta^* = \arg\min \mathcal{L}_{train}(\alpha, \theta) \tag{2}$$

where $\mathcal{L}_{train}$ is the objective in Eq. (1), θ denotes the concatenation of all classifier parameters $(\theta_E, \theta_c, \theta_1, ..., \theta_K)$, and $\mathcal{L}_{val}$ is loss of applying a trained model on a separate set of clean data. Note that $\theta^*(\alpha)$ denotes the dependency of θ^* on α after we train the classifier on a given LWN.

Analytically solving for the inner problem is typically infeasible. In this paper, we adopt the following one-step SGD update to approximate the optimal solution. As such the gradient for the meta-parameters α can be estimated as:

$$\nabla_\alpha \mathcal{L}_{val}(\theta - \eta \nabla_\theta \mathcal{L}_{train}(\alpha, \theta)) = -\eta \nabla^2_{\alpha,\theta} \mathcal{L}_{train}(\alpha, \theta) \nabla_{\theta'} \mathcal{L}_{val}(\theta') \tag{3}$$

$$\approx -\frac{\eta}{2\epsilon} \left[\nabla_\alpha \mathcal{L}_{train}(\alpha, \theta^+) - \nabla_\alpha \mathcal{L}_{train}(\alpha, \theta^-) \right] \tag{4}$$

where $\theta^\pm = \theta \pm \epsilon \nabla_{\theta'} \mathcal{L}_{val}(\theta')$, $\theta' = \theta - \eta \nabla_\theta \mathcal{L}_{train}(\alpha, \theta)$, ϵ is a small constant for finite difference and η is learning rate for SGD.

Since we leverage Multiple Weak Social Supervision, we term our method as MWSS. We adopt Adam with mini-batch training to learn the parameters. Algorithm 1 and Figure 3 outline the training procedure for MWSS.

3 Constructing Weak Labels from Social Engagements

In this section, we describe how to generate weak labels from users' social engagements that can be incorporated as weak sources in our model.

3.1 Dataset Description

We utilize one of the most comprehensive fake news detection benchmark datasets called FakeNewsNet [25]. The dataset is collected from two fact-checking websites: GossipCop[6] and PolitiFact[7] containing news contents with labels annotated by professional journalists and experts, along with social context information. News content includes meta attributes of the news (e.g., body text), whereas social context includes related users' social engagements on the news items (e.g., user comments in Twitter). Note that the number of news pieces in PolitiFact data is relatively small, and we enhance the dataset to obtain more weak labels. Specially, we use a news corpus spanning the time frame 01 January 2010 through 10 June 2019, from 13 news sources including mainstream British news outlets, such as BBC and Sky News, and English language versions of Russian news outlets such as RT and Sputnik, which are mostly related to political topics. To obtain the corresponding social engagements, we use a similar strategy as FakeNewsNet [25] to get tweets/comments, user profiles and user history tweets through the Twitter API and web crawling tools. For GossipCop data, we mask part of the annotated data and treat them as unlabeled data for generating weak labels from the social engagements.

[6] https://www.gossipcop.com/
[7] https://www.politifact.com/

3.2 Generating Weak Labels

Now, we introduce the labeling functions for generating weak labels from social media via statistical measures guided by computational social theories.

First, research shows user opinions towards fake news have more diverse sentiment polarity and less likely to be neutral [3]. So we measure the sentiment scores (using a widely used tool VADER [7]) for all the users sharing a piece of news, and then measure the variance of the sentiment scores by computing the standard deviation. We define the following weak labeling function:

> ***Sentiment-based:*** *If a news piece has a standard deviation of user sentiment scores greater than a threshold τ_1, then the news is weakly labeled as fake news.*

Second, social studies have theorized the correlation between the bias of news publishers and the veracity of news pieces [5]. Accordingly, we assume that news shared by users who are more biased are more likely be fake, and vice versa. Specifically, we adopt the method in [10] to measure user bias (scores) by exploiting users' interests over her historical tweets. The hypothesis is that users who are more left-leaning or right-leaning share similar interests with each other. Following the method in [10], we generate representative sets of people with known public bias, and then calculate bias scores based on how closely a query users' interests match with those representative users. We define the following weak labeling function:

> ***Bias-based:*** *If the mean value of users' absolute bias scores – sharing a piece of news – is greater than a threshold τ_2, then the news piece is weakly-labeled as fake news.*

Third, studies have shown that less credible users, such as malicious accounts or normal users who are vulnerable to fake news, are more likely to spread fake news [27]. To measure user credibility, we adopt the practical approach in [1]. The hypothesis is that less credible users are more likely to coordinate with each other and form big clusters, whereas more credible users are likely to form small clusters. We use the hierarchical clustering[8] to cluster users based on their meta-information on social media and take the reciprocal of the cluster size as the credibility score. Accordingly, we define the following weak labeling function:

> ***Credibility-based:*** *If a news piece has an average credibility score less than a threshold τ_3, then the news is weakly-labeled as fake news.*

To determine the proper thresholds for τ_1, τ_2, and τ_3, we vary the threshold values from $[0, 1]$ through binary search, and compare the resultant weak labels with the true labels from the training set of annotated clean data – later used to train our meta-learning model – on GossipCop, and choose the value that achieves the the best accuracy on the training set. We set the thresholds as $\tau_1 = 0.15$, $\tau_2 = 0.5$, and $\tau_3 = 0.125$. Due to the sparsity for Politifact labels, for simplicity, we use the same rules derived from the GossipCop data.

[8] https://bit.ly/2WGK6zE

Table 1: The statistics of the datasets. Clean refers to manually annotated instances, whereas the weak ones are obtained by using the weak labeling functions

Dataset	GossipCop	Politifact
# Clean positive	1,546	303
# Clean negative	1,546	303
# Sentiment-weak positive	1,894	3,067
# Sentiment-weak negative	4,568	1,037
# Bias-weak positive	2,587	2,484
# Bias-weak negative	3,875	1,620
# Credibility-weak positive	2,765	2,963
# Credibility-weak negative	3,697	1,141

Quality of Weak Labeling Functions We apply the aforementioned labeling functions and obtain the weakly labeled positive instances. We treat the news pieces discarded by the weak labeling functions as *negative* instances. The statistics are shown in Table 1. To assess the quality of these weakly-labeled instances, we compare the weak labels with the true labels on the annotated clean data in GossipCop – later used to train our meta-learning model. The accuracy of the weak labeling functions corresponding to Sentiment, Bias, and Credibility are 0.59, 0.74, 0.74, respectively. The F1-scores of these three weak labeling functions are 0.65, 0.64, 0.75. We observe that the accuracy of the labeling functions are significantly better than random (0.5) for binary classification indicating that the weak labeling functions are of acceptable quality.

4 Experiments

Now, we present the experiments to evaluate the effectiveness of MWSS. We aim to answer following evaluation questions: (1) **EQ1**: Can MWSS improve fake news classification performance by leveraging weak social supervision; (2) **EQ2**: How effective are the different sources of supervision for improving prediction performance; and (3) **EQ3**: How robust is MWSS on leveraging multiple sources?

4.1 Experimental Settings

Evaluation measures. We use F1 score and accuracy as the evaluation metrics. We randomly choose 15% of the clean instances for validation and 10% for testing. We fix the number of weak training samples and select the amount of clean training data based on the *clean data ratio* defined as: clean ratio = $\frac{\#\text{clean labeled samples}}{\#\text{clean labeled samples} + \#\text{weak labeled samples}}$. This allows us to investigate the contribution of clean vs. weakly labeled data in later experiments. All the clean datasets are balanced with positive and negative instances. We report results on the test set with the model parameters picked with the best validation accuracy. All runs are repeated for 3 times and the average is reported.

Base Encoders. We use the convolutional neural networks (CNN) [9] and RoBERTa-base, a robustly optimized BERT pre-training model [12] as the encoders for learning content representations. We truncate or pad the news text to 256 tokens, and for the CNN encoder we use pre-trained WordPiece embeddings from BERT to initialize the embedding layer. For each of the $K+1$ classification heads, we employ a two-layer MLP with 300 and 768 hidden units for both the CNN and RoBERTa encoders. The LWN contains a weak label embedding layer with dimension of 256, and a three-layer MLP with $(768, 768, 1)$ hidden units for each with a sigmoid as the final output function to produce a scalar weight between 0 and 1. We use binary cross-entropy as the loss function ℓ for MWSS[9].

Baselines and learning configurations. We consider the following settings:

(1) training only with limited amount of manually annotated **clean** data. Models include the following state-of-the-art early fake news detection methods:

- TCNN-URG [17]: This method exploits users' historical comments on news articles to learn to generate synthetic user comments. It uses a two-level CNN for prediction when user comments are not available for early detection.
- EANN [33]: This method utilizes an adversarial learning framework with an event-invariant discriminator and fake news detector. For a fair comparison, we only use the text CNN encoder.

(2) training only with **weakly** labeled data; and (3) training with both the **clean** and **weakly** labeled data as follows:

- Clean+Weak: In this setting, we simply merge both the clean and weak sets (essentially treating the weak labels to be as reliable as the clean ones) and use them together for training different encoders.
- L2R [23]: L2R is the state-of-the-art algorithm for learning to re-weight (L2R) examples for training models through a meta learning process.
- Snorkel [21]: It combines multiple labeling functions given their dependency structure by solving a matrix completion-style problem. We use the label generated by Snokel as the weak label and feed it to the classification models.
- MWSS: The proposed model for jointly learning with clean data and multi-sources of weak supervision for early fake news detection.

Most of the above baseline models are geared for single sources. In order to extend them to multiple sources, we evaluated several aggregation approaches, and found that taking the majority label as the final label achieved the best performance result. We also evaluate an advanced multiple weak label aggregation method – Snorkel [19] as the multi-source baseline. Note that our MWSS model, by design, aggregates information from multiple sources and does not require a separate aggregation function like the majority voting.

4.2 Effectiveness of Weak Supervision and Joint Learning

To answer **EQ1**, we compare the proposed framework MWSS with the representative methods introduced in Section 4.1 for fake news classification. We

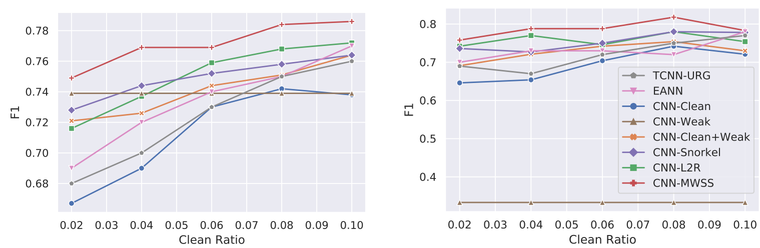
[9] All the data and code are available at: **this clickable link**

Table 2: Performance comparison for early fake news classification. *Clean* and *Weak* depict model performance leveraging only those subsets of the data; *Clean+Weak* is the union of both the sets.

Methods	GossipCop		PolitiFact	
	F1	Accuracy	F1	Accuracy
TCNN-URG (Clean)	0.76	0.74	0.77	0.78
EANN (Clean)	0.77	0.74	0.78	0.81
CNN (Clean)	0.74	0.73	0.72	0.72
CNN (Weak)	0.73	0.65	0.33	0.60
CNN (Clean+Weak)	0.76	0.74	0.73	0.72
CNN-*Snorkel* (Clean+Weak)	0.76	0.75	0.78	0.73
CNN-*L2R* (Clean+Weak)	0.77	0.74	0.79	0.78
CNN-MWSS (Clean+Weak)	**0.79**	**0.77**	**0.82**	**0.82**
RoBERTa (Clean)	0.77	0.76	0.78	0.77
RoBERTa (Weak)	0.74	0.74	0.33	0.60
RoBERTa (Clean+Weak)	0.80	0.79	0.73	0.73
RoBERTa-*Snorkel* (Clean+Weak)	0.76	0.74	0.78	0.77
RoBERTa-*L2R* (Clean+Weak)	0.78	0.75	0.81	0.82
RoBERTa-MWSS (Clean+Weak)	**0.80**	**0.80**	**0.82**	**0.82**

determine the model hyper-parameters with cross-validation. For example, we set parameters $learning\_rate \in \{10^{-3}, 5 \times 10^{-4}, 10^{-4}, 5 \times 10^{-5}\}$ and choose the one that achieves the best performance on the held-out validation set. From Table 2, we make the following observations:

– Training only on clean data achieves better performance than training only on the weakly labeled data consistently across all the datasets (clean > weak).
– Among methods that only use clean data with CNN encoders, we observe TCNN-URG and EANN to achieve relatively better performance than Clean consistently. This is because TCNN-URG utilizes user comments during training to capture additional information, while EANN considers the event information in news contents (TCNN-URG>CNN-clean, and EANN>CNN-clean).
– On incorporating weakly labeled data in addition to the annotated clean data, the classification performance improves compared to that using only the clean labels (or only the weak labels) on both datasets (demonstrated by clean+weak, L2R, Snorkel > clean > weak).
– On comparing two different encoder modules, we find that RoBERTa achieves much better performance in GossipCop compared to CNN, and has a similar performance in PolitiFact. The smaller size of the PolitiFact data results in variable performance for RoBERTa.
– For methods that leverage both the weak and clean data, L2R and Snorkel perform quite well. This is because L2R assigns weight to instances based on their contribution with a held-out validation set, whereas Snorkel leverages correlations across multi-source weak labeling functions to recover the label.



(a) GossipCop (b) PolitiFact

Fig. 4: F1 score with varying clean data ratio from 0.02 to 0.1 with CNN-MWSS. The trend is the similar with RoBERTa encoder (best visualized in color).

– In general, our model MWSS achieves the best performance. We observe that MWSS $>$ L2R and Snorkel on both the datasets. This demonstrates the importance of treating weak labels differently from the clean labels with a joint encoder for learning shared representation, separate MLPs for learning source-specific mapping functions, and learning to re-weight instances via LWN. To understand the contribution of the above model components, we perform an ablation study in the following section.

4.3 Impact of the Ratio of Clean to Weakly Labeled Data on Classification Performance

To answer **EQ2**, we explore how the performance of MWSS changes with the clean ratio. We set the clean ratio to vary in $\{0.02, 0.04, 0.06, 0.08, 0.1\}$. To have a consistent setting, we fix the number of weakly labeled instances and change the number of clean labeled instances accordingly. In practise, we have abundant weak labels from the heuristic weak labeling functions. The objective here is to figure out how much clean labels to add in order to boost the overall model performance. Figure 4 shows the results. We make the following observations:

– With increasing values of clean ratio, the performance increases for all methods (except *Weak* which uses a fixed amount of weakly labeled data). This shows that increasing amount of reliable clean labels obviously helps the models.
– For different clean ratio configurations, MWSS achieves the best performance compared to other baselines, i.e., MWSS $>$ L2R and Snorkel. This shows that MWSS can more effectively utilize the clean and weak labels via its multi-source learning and re-weighting framework.
– We observe that the methods using Clean+Weak labels where we treat the weak labels to be as reliable as clean ones may not necessarily perform better than using only clean labels. This shows that simply merging the clean and weak sources of supervision without accounting for their reliability may not improve the prediction performance.

Table 3: F1/Accuracy on training MWSS on different weak sources with clean data.

Dataset	Sentiment	Bias	Credibility	All Sources
GossipCop	0.75/0.69	0.78/0.75	0.77/0.73	0.79/0.77
PolitiFact	0.75/0.75	0.77/0.77	0.75/0.73	0.78/0.75

Table 4: F1/Accuracy result of ablation study on modeling source-specific MLPs with different clean ratio (C-Ratio). "SH" denotes a single shared MLP and "MH" denotes multiple source-specific ones.

Model	C-Ratio	L2R	LWN
SH	0.02	0.72/0.68	0.73/0.72
	0.10	0.77/0.74	0.77/0.73
MH	0.02	0.73/0.71	0.75/0.71
	0.10	0.78/0.76	0.79/0.77

Fig. 5: Label weight density distribution among weak and clean instances in GossipCop. The mean of the label weight for *Credibility*, *Sentiment*, *Bias* and *Clean* are 0.86, 0.85, 0.86 and 0.87 respectively.

4.4 Parameter Analysis

Impact of source-specific mapping functions: In this experiment, we want to study the impact of modeling separate MLPs for source-specific mapping functions (modeled by f_{θ_k} in Equation 1) in LWN and L2R as opposed to replacing them with a single shared MLP (i.e. $f_{\theta_k} = f_\theta \ \forall k$) across multiple sources. From Table 4, we observe that MWSS and L2R both work better with multiple source-specific MLPs as opposed to a single shared MLP by better capturing source-specific mapping functions from instances to corresponding weak labels. We also observe MWSS to perform better than L2R for the respective MLP configurations – demonstrating the effectiveness of our re-weighting module.

Impact of different weak sources: To study the impact of multi-source supervision, we train MWSS separately with individual weak sources of data along with clean annotated instances with a clean ratio of 0.1. From Table 3, we observe that training MWSS with multiple weak sources achieves better performance compared to that of a single weak source – indicating complementary information from different weak sources help the model. To test whether MWSS can capture the quality of each source, we visualize the label weight distribution for each weak source and clean dataset in Figure 5. From the weight distribution, we also observe the weight of the sentiment-source (referred as *Sentiment*) to be less than that of other sources. In addition, although the LWN is not directly trained on clean samples, it still assigns the largest weight to the clean source.

These demonstrate that our model not only learns the importance of different instances but also learns the importance of the corresponding source.

5 Related Work

Fake News Detection: Fake news detection methods mainly focus on using *news contents* and with information from *social engagements* [27]. Content-based approaches exploit feature engineering or latent features to capture deception cues [16,24]. For social context based approaches, features are mainly extracted from users, posts and networks. User-based features are extracted from user profiles to measure their characteristics [2]. Post-based features represent users' responses in term of stance, topics, or credibility [2,8]. Network-based features are extracted by constructing specific networks. Recently, deep learning models are applied to learn the temporal and linguistic representation of news [17,33]. Wang *et al.* proposed an adversarial learning framework with an event-invariant discriminator and fake news detector [33]. Qian *et al.* exploited users' historical comments on news articles to learn to generate synthetic user comments for early fake news detection [17].

Learning with Weak Supervision: Most machine learning models rely on quality labeled data to achieve good performance where the presence of label noise or adversarial noise cause a dramatic performance drop [22]. Therefore, learning with noisy labels has been of great interest for various tasks [26,34]. Existing works attempt to rectify the weak labels by incorporating a loss correction mechanism [30,14,11]. Patrini *et al.* [14] utilize the loss correction mechanism to estimate a label corruption matrix without making use of clean labels. Other works consider the scenario where a small set of clean labels are available [6,23,35]. For example, Zheng *et al.* propose a meta label correction approach using a meta model which provides reliable labels for the main models to learn. Recent works also consider that weak signals are available from multiple sources [18,31,20] and consider the redundancy and consistency across labels. In contrast, the weak signals in our work are derived from user engagements and we do not make any assumptions about the structure of the label noise.

6 Conclusions and Future Work

In this paper, we develop techniques for early fake news detection leveraging weak social supervision signals from multiple sources. Our end-to-end framework MWSS is powered by meta learning with a Label Weighting Network (LWN) to capture the varying importance weights of such weak supervision signals from multiple sources during training. Extensive experiments in real-world datasets show MWSS to outperform state-of-the-art baselines without using any user engagements at prediction time. As future work, we want to explore other techniques like label correction methods to obtain high quality weak labels to further improve our models. In addition, we can extend our framework to consider other types of weak social supervision signals from social networks leveraging temporal footprints of the claims and engagements.

Acknowledgments

This work is, in part, supported by Global Security Initiative (GSI) at ASU and by NSF grant # 1614576.

References

1. Mohammad-Ali Abbasi and Huan Liu. Measuring user credibility in social media. In *SBP*, 2013.
2. Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *WWW*, 2011.
3. Limeng Cui and Suhang Wang Dongwon Lee. Same: Sentiment-aware multi-modal embedding for detecting fake news. 2019.
4. Liang Ge, Jing Gao, Xiaoyi Li, and Aidong Zhang. Multi-source deep learning for information trustworthiness estimation. In *KDD*. ACM, 2013.
5. Matthew Gentzkow, Jesse M Shapiro, and Daniel F Stone. Media bias in the marketplace: Theory. In *Handbook of media economics*, volume 1, pages 623–645. Elsevier, 2015.
6. Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, 2018.
7. Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*, 2014.
8. Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *AAAI*, 2016.
9. Yoon Kim. Convolutional neural networks for sentence classification, 2014.
10. Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. Quantifying search bias: Investigating sources of bias for political searches in social media. In *CSCW*. ACM, 2017.
11. Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *ICCV*, 2017.
12. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.
13. Wanli Ouyang, Xiao Chu, and Xiaogang Wang. Multi-source deep learning for human pose estimation. In *CVPR*. IEEE, 2014.
14. Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 2017.
15. James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.
16. Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.

17. Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. Neural user response generator: Fake news detection with collective user intelligence. In *IJCAI*, 2018.
18. Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017.
19. Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, pages 1–22, 2019.
20. Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training complex models with multi-task weak supervision. *arXiv preprint arXiv:1810.02840*, 2018.
21. Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training complex models with multi-task weak supervision, 2018.
22. Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.
23. Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. *arXiv preprint arXiv:1803.09050*, 2018.
24. Victoria L Rubin and Tatiana Lukoianova. Truth and deception at the rhetorical structure level. *Journal of the Association for Information Science and Technology*, 66(5):905–917, 2015.
25. Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.
26. Kai Shu, Subhabrata Mukherjee, Guoqing Zheng, Ahmed Hassan Awadallah, Milad Shokouhi, and Susan Dumais. Learning with weak supervision for email intent detection. *SIGIR*, 2020.
27. Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *KDD exploration newsletter*, 2017.
28. Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *WSDM*, 2019.
29. Russell Stewart and Stefano Ermon. Label-free supervision of neural networks with physics and domain knowledge. In *AAAI*, 2017.
30. Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.
31. Paroma Varma, Frederic Sala, Ann He, Alexander Ratner, and Christopher Ré. Learning dependency structures for weak supervision models. *ICML*, 2019.
32. Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 2018.
33. Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *CIKM*, 2018.
34. Zhen-Yu Zhang, Peng Zhao, Yuan Jiang, and Zhi-Hua Zhou. Learning from incomplete and inaccurate supervision. In *KDD*, 2019.
35. Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for learning with weak supervision. *arXiv preprint arXiv:1911.03809*, 2019.