Evidence-Aware Multilingual Fake News Detection

HICHAM HAMMOUCHI, (Member, IEEE), AND MOUNIR GHOGHO, (Fellow, IEEE)
TICLab, College of Engineering and Architecture, International University of Rabat, Salé 11100, Morocco
Corresponding author: Hicham Hammouchi (hicham.hammouchi@uir.ac.ma)

This work was supported by the e-Covid19 Project through the Centre National de la Recherche Scientifique et Technique (CNRST).

ABSTRACT Due to the rapid growth of the Internet and the subsequent rise of social media users, sharing information has become more flexible than ever before. This unrestricted freedom has also led to an increase in fake news. During the Covid-19 outbreak, fake news spread globally, negatively affecting authorities’ decisions and the health of individuals. As a result, governments, media agencies, and academics have established fact-checking units and developed automatic detection systems. Research approaches to verify the veracity of news focused largely on writing styles, propagation patterns, and building knowledge bases that serve as a reference for fact checking. However, little work has been done to assess the credibility of the source of the claim to be checked. This paper proposes a general framework for detecting fake news that uses external evidence to verify the veracity of online news in a multilingual setting. Search results from Google are used as evidence, and the claim is cross-checked with the top five search results. Additionally, we associate a vector of credibility scores with each evidence source based on the domain name and website reputation metrics. All of these components are combined to derive better predictions of the veracity of claims. Further, we analyze the claim-evidence entailment relationship and select supporting and refuting evidence to cross-check with the claim. The approach without selection components yields better detection performance. In this work, we consider as a case study Covid-19 related news. Our framework achieves an F1-score of 0.85 and 0.97 in distinguishing fake from true news on XFact and Constraint datasets respectively. With the achieved results, the proposed framework present a promising automatic fact checker for both early and late detection.

INDEX TERMS Information extraction, fake news detection, classification, evidence-aware, source credibility.

I. INTRODUCTION Social media platforms such as Facebook, Twitter, and YouTube, as well as news agencies have taken steps to eliminate false information from the online space and to provide reliable information. They have addressed the issue mainly by establishing fact-checking teams that monitor and verify claims. An undertaking of this magnitude requires high levels of sustainability and scalability. As a matter of fact, many organizations that established fact-checking services have discontinued them. Another approach to the detection of fake news involves the use of machine learning techniques to classify claims automatically. Many approaches have been adopted, including analyzing the writing style and the propagation patterns of fake news, as well as building knowledge bases to use as a reference to verify emerging claims. Fake news detection methods can be classified into four categories [1]: 1) knowledge-based, 2) style-based, 3) propagation-based and 4) source-based. There has not been enough attention given to this latter category in the literature, even though a reliable method for measuring the integrity of the source would greatly benefit the detection system. Integrating source-based approaches with other approaches within a comprehensive system could be more effective at identifying false news.

We address the problem of automatic fact-checking of online news articles by leveraging existing evidence and assessing the credibility of the source of the evidence.
We emphasize the importance of verifying claims through the inclusion of evidence and the assessment of the credibility of the sources providing the evidence. By combining these elements with the claim, the detection model could better distinguish between fake news and real news. To verify claims and evidence regardless of their language, we encode them using Bert multilingual and feed the resulting representations into the Long Short Term Memory (LSTM) layers. In this way, we enhance the model’s ability to function in different contexts independent of the language used. This paper examines Covid-19 as a case study to evaluate the proposed framework. Our contributions are as follows:

1) Motivated by a lack of work on source credibility, we propose an approach based on website content and domain name reputation metrics for assessing the credibility of sources;
2) To improve the generalizability of the fact-checking system, we introduce a multilingual architecture composed of Bert representations and LSTM models to cross-check claims with their evidence, as well as a credibility score assigned to the source of the evidence;
3) We analyze the claim-evidence entailment as a evidence selection mechanism to distinguish supporting and refuting evidence;
4) We curate a new sample of Covid-19-related news in Morocco for the early detection.

The remainder of the paper is organized as follows. In section II, we review previous research on fake news detection. We present the proposed method in detail in section III. In section IV, we present the conducted experiments and discuss the results. Section V concludes the papers.

II. RELATED WORK

Due to the popularity of social media, information dissemination has become more open, which has resulted in the spread of inaccurate and false information. Initial attempts to solve the problem of fake news detection relied primarily on manual fact checking by experts and with little automated analysis of textual features and style [2], [3], [4], [5], [6]. While manual fact-checking is still a key component in combating fake news, it should be reinforced with automated methods as fake news and misinformation continue to proliferate and become more sophisticated. Researchers have also examined the reasons behind fake news and categorized them into confirmation bias, overacceptance of weak claims, and lack of analytic ability [7], [8].

Building an automatic system for the detection of fake news requires the development of a representative and comprehensive dataset, particularly for an emerging field such as Covid-19. In this context, considerable efforts have been made to create datasets based primarily on tweets from Twitter. For example, to gauge Moroccans’ emotional response to the pandemic and government decisions, Ghanem et al. [9] developed a real-time infoveillance platform and collected comments related to Covid-19 from the most popular social media platforms in Morocco, including Twitter, Facebook, and Youtube, as well as two popular news sites. The dataset contained over 747K comments expressed in Moroccan dialect and modern standard Arabic throughout 2020. Haouari et al. introduced Arcov19 [10] and Arcov19-Rumors [11] as two datasets containing Arabic fact checked claims along with related tweets and their propagation (i.e. retweets and replies networks). Paka et al. [12] introduced the CTF dataset, a collection of Covid-19 related English tweets. In addition, Constraint dataset [13], which contains fake and real news related to Covid-19, was collected and published as part of the Association for the Advancement of Artificial Intelligence conference 2021. Gupta et al. [14] collected fact-checked news from a variety of fact-checking websites covering a wide range of topics.

Methods of detecting fake news include analyzing text style and grammar features, understanding propagation patterns, and verifying against a knowledge base to assess the source’s credibility. For the textual analysis, previous work has either developed a set of features to alleviate the differences between false and true claims, or has used established methods for linguistic analysis such as LIWC [5]. In this sense, Choudhary et al. [2] defined a set of features characterizing syntax, sentiment, grammatical structure and readability of the text. The syntax features include, for example, the number of upper case and stop words, the sentiment features cover polarity and subjectivity, grammatical features include how the sentence is broken down into noun, verb, adjective, etc., and for the readability feature, they analyzed a predefined score to determine the level of readability. A neural network was then used to classify these features. Similarly, Singh et al. [4] studied the coherence of fake news and the differences between legitimate and fake news in terms of the length of the news, the structure and the coherence. They found that fake news tend to be shorter and less coherent than real news. Others have used a bi-modal approach combining text and images [15], [16]. Ozbay and Alatas [17] on the other hand, approached the fake news detection problem as an optimization problem to search for the best model that represents best the two classes. They used Grey Wolf Optimization and Salp Swarm Optimization algorithms and achieved decent results on the provided datasets outperforming supervised machine learning models.

Another approach that has been considered is to study the propagation of fake news through social networks. Cheng et al. [18] addressed the inference of causal patterns in the propagation networks of fake news in order to understand why people share certain news. They examined the relationship between user profiles and the likelihood of sharing fake news. Other previous work explored the relationship between news spreaders and user profiles. For instance, Shu et al. [19] highlighted the role social context plays in disseminating fake news. They proposed TriFN to combine news content, user engagement on social media, as well as the interrelationship between user-news and publisher-news to feed them to a classifier. Zhou and Zafarani [20] were inspired by psychological
theories and developed a network-based approach that leverages spreading patterns at node, ego, triad, community, and network levels. Using social psychology theories, they examined a number of inherent features, such as user susceptibility, user influence, user engagement, and network characteristics. Another work done by Lu and Li [21] proposed a graph-aware co-attention architecture that model the interaction between retweeters as a graph and model the propagation using ConvNet and GRU networks to predict whether a tweet is fake or real. Shi et al. [22] on the other hand analyzed the structure of the network by building micro-level and macro-level propagation networks representing replies and retweets, respectively. They conducted a structural and temporal analysis to look at the characteristics of the network in terms of depth, outdegree as well as the temporal differences between retweets and replies, and between replies in a cascade. Overall, the researchers found that micro-level features performed better, and temporal features performed better compared to structural features. Vo and Lee [23] proposed an evidence-aware architecture combining claims with related articles using multi-head attention. Similarly, Dou et al. [24] proposed an architecture that incorporates news content, user historical posts informing on the user engagement, as well as the propagation tree of the news encoding the sharing cascade of a claim.

We argue that a critical missing piece of the study be the explainability of such detection, i.e., why a particular piece of news is detected as fake. In this paper, therefore, we study the explainable detection of fake news. We develop a sentence-comment co-attention sub-network to exploit both news contents and user comments to jointly capture explainable top-k check-worthy sentences and user comments for fake news detection.

Knowledge and evidence-based methods seek to verify the veracity of a claim by relying on prior knowledge. A recent study by Shu et al. [25] addressed the problem of explainability of fake news detection systems. They proposed a co-attention network to explain why a news item is fake based on related comments posted on social media. Hu et al. [26] focused on detecting fake news by comparing claims against an external knowledge via a graph neural model. They built a directed heterogeneous document graph to represent topics and entities corresponding to each news item. Similarly, Vijalli et al. [27] proposed a two-stage transformer that generates a set of candidate explanations for a given claim, then uses the selected explanations to compute the textual entailment between the claim and retrieved facts. Baris and Boukhers [28] used knowledge about previously published fake news and information about the source to detect fake news related to covid. In contrast, Wang et al. [29] looked at the fake news detection problem as a weak supervision problem by using reinforcement learning to annotate unlabeled articles. They used users’ reports as a weak supervision to select quality reports via a reinforced selector and annotate the unlabeled items to finally train a fake news detector.

An important part of verifying the veracity of a claim is evaluating its credibility. In [30], the authors examined the likelihood of a social media user spreading fake news based on their profile features. Zhang et al. [31] studied information about published articles and their creators in social media and examined the problem as a credibility inference. Esteves et al. [32] developed a model for categorizing websites into binary and multi-class categories. They extracted a set of textual features and look at 15 features such as domain name authority, HTML2Seq consisting of bag of tags, etc. In a similar vein, Zhou et al. [33] examined the credibility of news websites through a set of features describing their behavior. Among these features are whether the website repeatedly publishes false content, avoids deceptive headlines, etc. Credibility of online news can be measured from the source and content perspectives [34]. Credibility of a source is mainly determined by the credibility of the author.

In this work, the credibility score of the evidence source is measured using the reputation of the corresponding domain name and website pages. By measuring this score independently of the content of the claim, we enrich the representation of the evidence.

III. THE PROPOSED METHOD

Our approach is evidence-oriented and relies on verification of claims by examining related evidence and assessing the credibility of the source. The evidence can be presented in the form of articles in news publications, and facts supplied by governmental agencies and international organizations that corroborate the claim or refute it. Fact checking emerging claims by taking into account the aforementioned evidence and source credibility would enrich the representation of the input claim, although the process is not straightforward. Further, assuming all evidence is equal in terms of the credibility of the information they convey can be misleading and hurt rather than help a model’s performance. In order to alleviate this problem, we utilize search engine optimization (SEO) metrics to evaluate the credibility of the sources in terms of domain authority and website reputation. This is motivated by the fact that information coming from highly credible sources would contain the useful elements to refute or validate a claim. In fact, highly credible sources have high credibility scores (i.e., high domain authority scores). We focus on these elements to fact check emerging claims, while the study of the propagation of news on social networks is also valid and significant for this task.

This section describes the entire process of data collection, processing, encoding, and model development for fact-checking online news.

A. INPUT DATA

As Covid-19 pandemic emerged unexpectedly, the lack of data was a major challenge facing researchers. There has been considerable efforts in creating datasets either from social media or news websites and annotating them manually for the purpose of training machine learning models in a supervised
manner. It is still difficult to find labeled datasets that contain fake news associated with Covid-19. In this work, we use XFact [14], one of the existing datasets about fake news containing fact checked news in 17 languages. As it contains news from many different sources, we have applied a filter to keep only Covid-19-related news. We also use Constraint dataset [35], a labeled dataset collected from social media platforms such as Twitter, Facebook, and Instagram, which contains Covid-19-related claims in English curated for the fake news detection task. Further, similar to the work [9] done on Moroccan comments on social media to manage the Covid-19 pandemic, we have collected additional news related to Covid-19 in Morocco to serve as a test sample and to further evaluate the proposed detection model. We collect French fact checked news from le360 website. This small sample will not be used during the training phase, but only at test time to evaluate the proposed framework on new data related to Covid-19.

B. DATA PROCESSING

1) FILTERING COVID-19-RELATED NEWS

The XFact dataset contains news about different topics. So, in order to extract only Covid-19 related claims, we first run the Latent Dirichlet Allocation (LDA) algorithm, which is an unsupervised topic modeling method, to find groups of topics in the dataset and the main keywords corresponding to Covid-19 news. Using these keywords, we filtered

the original XFact dataset to retain only news related to Covid-19 published between January 2020 and November 2020. In addition to the original claim, we translate all claims to English using Google translator. Based on the translated claims we keep only claims containing one or several of the

| Claim | Stanford study found that masks are ineffective against Covid19 |
|---|---|
| Date | 2021-04-07 |
| Country | Brazil |
| Language | English |
| Label | False |
| Evidence #1 | Study falsely claiming masks are harmful isn’t Stanford’s |
| Evidence #1 source | uioday.com |
| Evidence #2 | Surgical masks reduce COVID-19 spread, large-scale study shows med.stanford.edu |
| Evidence #2 source | |

| Claim [Translated] | Morocco returns to ground zero by applying general quarantine and closing all cities |
|---|---|
| Date | 2020-08-06 |
| Country | Morocco |
| Language | Arabic |
| Label | False |
| Evidence #1 [Translated] | The increase of Corona cases scares Moroccans who do not want to go back to zero. A new situation for Moroccans after the decision to lift the quarantine came into effect; Officially, the majority of Moroccan cities have returned to normal following the lifting of the quarantine hexpress.com |
| Evidence #1 source | |

1https://fr.le360.ma/
FIGURE 2. The most common words in the dataset and their relations in terms of co-occurrence.

following keywords: ['covid', 'virus', 'kovid', 'pandemic', 'epidemic', 'quarantine', 'vaccine', 'outbreak', 'corona', 'sars', 'curfew', 'lockdown', 'korona']. By selecting only claims related to Covid-19, we end up with 1003 claims, which are divided into 869 fake claims and 134 true claims.

2) NEWS ANALYSIS
The analysis of claims text translated to English, illustrated in fig. 2, shows that the most common words in the dataset are 'Corona', 'virus', 'Chinese', 'infected', 'vaccine', etc. with 'Corona' and 'virus' being the two words that appear the most together, which is natural. Two unexpected words, 'Bill' and 'Gates', refer to Bill Gates, who was wrongly accused of planning the pandemic on social media in 2020. For Moroccan claims, on top of ‘corona’ and ‘virus’, the most frequent words characterizing false claims are ‘issued’, ‘published’, ‘protect’ and ‘ministry’, which may have been used to persuade people that the claim is legitimate; see 3(a). Figure 3(b) shows that the main words in the true news are ‘protective’ and ‘masks’, which refer to the government’s instructions regarding wearing masks and adhering to protective measures. Further, ‘financial’, ‘fund’ were common and may refer to the government’s financial assistance to certain categories of citizens.

C. EVIDENCE COLLECTION
In order to verify the veracity of a claim, a human fact-checker should first look at evidence and check whether it supports or contradicts the claim. Further, the fact-checker selects from the available evidences those that look most reliable. A claim, if it is supported by a news item published on a reliable website, is likely to be true, and vice versa. In an attempt to simulate this process, we obtain the available evidences from the Google search results of the claim. We consider the top five articles returned by the search engine. We use the retrieved evidence articles to further investigate the claim since they may contain additional information that may be useful to validate or debunk the claim. Table 2 shows an example of retrieved evidence which contains useful pieces of information related to the claim.

In the present work, we use the XFact dataset which in fact contains the corresponding evidences (i.e. top five Google search results) to each claim. For the Moroccan sample of news and the Constraint dataset, we follow the same process to enrich the dataset with the related evidences. Table 1 shows an example of two claims in the dataset.

1) EVIDENCE RETRIEVAL
In order to find evidences related to a given claim, we search for the original claim using Google search and we retrieve the top five returned results as in [14]. We exclude FAQs if any and keep the title of the article and its link. For the title, we perform a basic preprocessing to remove any indication of the news media organization. For example, in the following search result: “Morocco kicks off coronavirus vaccination drive | Africanews”, we remove the substring after the vertical slash ‘|’, the news agency name “Africanews” and any other indication of false or true news ['False/Fake','True'] to keep only the title, and prevent the model from overfitting a specific term in the prediction.

2) CREDIBILITY MEASURE
The idea here is that websites that provide evidence should be trusted sources and have good quality backlinks as references to their articles. Evaluating the credibility of a source would add valuable information to the classification process. To determine the credibility of evidence source, we rely on
Search Engine Optimization (SEO) metrics to get domain name authority and website pages reputation. Metrics about search engine optimization are provided by a variety of sources, particularly by Mozilla (MOZ), which is one of the most reliable SEO providers that offers tools to measure the reputation of websites. We parse the retrieved links for each evidence article in our dataset to get the domain name, then using "Website SEO Checker". The Website SEO Checker aggregates SEO metrics from different sources to provide a more holistic view of a domain name, as the example in Table 3. In the present study, we consider the following metrics:

- Domain name authority (DA): a ranking score that determines how a domain ranks in search engines. It is calculated based on several metrics such as quality content and social signals from social media [36]. It takes a value between 1 and 100, and the higher the score the better the domain authority. Values between 40 and 50 are considered as average; values between 50 and 60 as good, and values above 60 are considered excellent.
- Page authority (PA): this is the value a search engine assigns to a Web page depending on its content quality and directly impacts the domain authority [36].
- Percentage of quality Backlinks (PQ): this represents the relevance of a webpage as determined by the quality of the inbound links referencing it [37].
- MozTrust score (MT): based on a 1:10 scale, this is a measure of the trustworthiness of pages and traffic they receive from trusted websites.
- Spam score (SS): this measures how the webpage content is similar to spamming pages.
- Off-page SEO score (OS): this represents how other websites perceive a webpage, which can be based on posted reviews.

### D. THE EMBEDDING LAYER

The embedding layer ensures that words in the input text are mapped to a latent vector space with a richer representation that preserves syntactic and semantic information.

### E. LSTM LAYER

In our proposed architecture, the LSTM layer takes as inputs the embeddings generated by mBert for each word and outputs a learned representation of the input text. The resulting representation is then combined with the credibility scores and fed to the classifier.

The LSTM is a family of Recurrent Neural Network (RNN) which can be described using deterministic transitions from the previous hidden state to the current state. Hidden states are denoted by \( h = (h_1, \cdots, h_T) \) and the output vector \( y = (y_1, \cdots, y_T) \) is obtained by iterating the following equations from \( t = 1 \) to \( T \) as shown in fig. 4 and given in [38]:

\[
h_t = \mathcal{H}(W_{ih}x_t + W_{hh}h_{t-1} + b_h) \quad (1)
\]

\[
y_t = W_{hy}h_t + b_y \quad (2)
\]

2https://weblizard.com/domain-authority-checker we get the available metrics about the domain

3https://huggingface.co/bert-base-multilingual-cased

---

**TABLE 3. An example of domain authority metrics: the first two domains are reliable because they have high domain authority and a high percentage of quality backlinks. The last two domains are known for publishing unverified news.**

| Domain             | DA | PA | PQ      | MT | SS | OS |
|--------------------|----|----|---------|----|----|----|
| bbc.com            | 95 | 75 | 90%     | 8  | 1% | 88%|
| unicef.org         | 92 | 65 | 93%     | 7  | 1% | 84%|
| burrardstreetjournal.com | 46 | 36 | 42%     | 4  | 1% | 53%|
| habapress.com      | 32 | 46 | 52%     | 5  | 28%| 50%|
where $H$ is either a logistic sigmoid or a tanh function, $W$ refers to the weight matrices connecting layers, and $b$ denotes the bias vectors.

The Long Short Term Memory (LSTM) has an advanced mechanism that allows it to memorize not only information from one previous layer but for an extended number of timesteps.

In the case of natural language processing, it is important to learn the sentence structure from both directions. The Bidirectional LSTM (Bi-LSTM) with two separate hidden layers allows to compute the forward pass $\vec{h}$ and the backward pass $\vec{h}$ and the output vector $y$ by iterating the backward layer from $t = T$ to 1, the forward layer from $t = 1$ to $T$ and then updating the output layer, as illustrated in the following equation and fig. 5:

$$y_t = W_y \vec{h}_t + W_y \vec{h}_t + b_y$$

### IV. EXPERIMENTS AND DISCUSSION

#### A. CLAIM-EVIDENCE ENTAILMENT CHECK

Before learning the structures of claims and their related evidence, we address the claim classification through claim-evidence entailment classification. We check whether we need a model for classification or we just need to study the entailment between the claim and its related evidence to predict the claim’s class. To do so, we introduce a Bert model to classify the relationship between claim and evidence into “entailment” (i.e. evidence supports the claim), “neutral” (i.e. no obvious relationship) and “contradiction” (i.e. evidence refutes the claim). The claim’s class is derived from the aforementioned classification. The claim is fake if the majority of predicted classes are “neutral” or “contradiction”, otherwise, the claim is true. For this task, we use a Bert model curated for natural language inference (NLI). Specifically, We use DeBERTa-v3-mnli-fever-anli-ling-wanli model [39], a version of DeBertav3 model [40] trained on four NLI and fact extraction and verification datasets, namely, MultiNLI, FEVER-NLI, ANLI, LingNLI and WANLI.

Based on the claim-evidence entailment relationship, we update our framework’s architecture and conduct additional experiment to consider only supporting and refuting evidence as depicted in Fig. 6. The average of the embeddings of each evidence type (i.e. support and refute) and concatenate them with the representation of the claim.

#### B. EVIDENCE-AWARE MULTILINGUAL FRAMEWORK

The main idea of our work is to fact-check claims based on available evidence and their credibility scores. The credibility component provides additional information on the trustworthiness of the source providing the evidence. For instance, evidence coming from a trusted and official organization will have a higher credibility score, which is accounted for in the training process. The credibility scores can be categorized into excellent, good, average and poor which can be translated to the level of credibility of the evidence source. In Table 5, we show the distribution of the credibility levels in each class. We build our architecture by processing in parallel the claim and its translated copy, evidence and source evidence
components. We learn the representations of the textual inputs through LSTM models and concatenate the scalar inputs with the resulting representation. All representation are concatenated and fed to a final classifier for claim validation or debunk.

C. TRAINING SETUP
To train our model, we consider the parallel setting depicted in figure 1 and we conduct our experiments on a workstation with an Intel i9-10980XE CPU, 128GB RAM and a Titan-Xp (12GB VRAM) GPU. We conduct all our experiments in the same conditions with the following parameters:

- # lstm cells = 100
- dense layer (size = 128, activation = leaky relu)
- optimizer = Adam(learning rate = 0.003), RectifiedAdam(learning rate = 0.007) with a reducing factor of 0.0005 for the learning rate
- epochs = 50, with early stopping

The dataset is divided into three sets for training, validation, and testing as shown in table 4. For the evaluation metrics, we consider the accuracy and F1 score for each class, defined by the following equations:

\[
F1-score_{class=c} = \frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c} \quad (4)
\]

\[
Accuracy_{class=c} = \frac{\# \text{ of correctly classified examples}}{\text{The total # of examples in class } 'c'} \quad (5)
\]

D. RESULTS
The proposed framework is composed of several components including the evidence integration and evidence source credibility scores. Therefore, to evaluate the detection performance, we conduct various experiments and ablations. These experiments are to evaluate whether the added components have added value to the system. Also, as a baseline, we consider the detection using only the claim at hand and learn the underlying language structure to distinguish fake from real news.

Table 6 shows the results of the fake news detection task on the test set with different settings. We report the accuracy score (percentage of correctly classified examples), the F1 score for each class, and the weighted F1 score. We consider the baseline model which is based only on the claim text. In addition, we test the proposed model on two completely new sets of examples relating to Covid-19 in Morocco published in Arabic and French. When using the claim text only, the proposed model achieves an F1 score of 0.79. When considering the claim, the evidence and credibility metrics, the F1 score increases up to 0.85. Our method thus outperforms prior work [14] which was based on fine-tuning multilingual Bert (mBert) using the XFact dataset. It is also worth pointing out that even when using the claim text only, our model outperforms that [14] when applied to the Covid-19 samples that we have investigated. The reason is that our architecture takes as input embeddings of the text and further learns the structure of the sentence through the LSTM layer. Moreover, the BiLSTM model achieves a slight improvement over the ordinary LSTM model. This can be attributed to the fact that the BiLSTM learns a richer representation through the forward and backward passes.

On the Constraint dataset, we achieve an f1-score of 0.97 on the test set, thus significantly outperforming the baseline and approaching the method using an ensemble of Roberta models and the COVID-Twitter-BERT (CT-BERT) model fine-tuned on Covid-19 related tweets.

On the claim-evidence entailment study, we achieve lower performance on both XFact and Constraint datasets as shown in Table 7. The separation between supporting and refuting evidence works as a selection process to elect only non-neutral evidence, which hurts the overall performance of the framework. This means that the proposed framework benefits from the enriched context and the selection is done as an underlying process in the learning.
TABLE 6. Fake news detection F1 score and accuracy results in different settings in comparison with mBert model in [14]. Evid denotes evidences and Cred denotes credibility metrics. 1-class is the true class, 0-class is the fake class, and Acc is the accuracy.

| Model                  | (1-class) Acc | (0-class) Acc | Avg. Acc | (1-class) F1 | (0-class) F1 | W. avg. F1 |
|------------------------|---------------|---------------|----------|--------------|--------------|------------|
| Claim only             | 6%            | 92%           | 78.30%   | 0.12         | 0.92         | 0.79       |
| Claim + Evid           | 21%           | 96%           | 84.06%   | 0.32         | 0.92         | 0.82       |
| Claim + Evid + Cred    | 35%           | 97%           | 87.12%   | 0.39         | 0.93         | 0.84       |
| Claim + Evid + Cred (BiLSTM) | 40%       | 98%           | 88.76%   | 0.43         | 0.93         | 0.85       |
| Claim + Evid (mBert fine-tuned [14]) | -          | -             | -        | -            | -            | 0.70       |

Moroccan Fact checked news from www.le600.ma

| Model                  | (1-class) Acc | (0-class) Acc | Avg. Acc | (1-class) F1 | (0-class) F1 | W. avg. F1 |
|------------------------|---------------|---------------|----------|--------------|--------------|------------|
| Claim + Evid + Cred (BiLSTM+FC) | 85%          | 37.5%         | 54.6%    | 0.79         | 0.44         | 0.64       |
| Claim + Evid + Cred (BiLSTM+XGB) | 85%          | 35.5%         | 53.32%   | 0.76         | 0.44         | 0.64       |
| Claim + Evid (w/o Cred) | 95%           | 22.3%         | 48.6%    | 0.79         | 0.35         | 0.62       |
| Claim only (w/o Evid) FC | 100%          | 0%            | 36%      | 0.77         | 0.0          | 0.48       |
| Claim only (w/o Cred, w/o Evid) XGB | 88%          | 10%           | 38.08%   | 0.73         | 0.15         | 0.51       |

E. DISCUSSION

Since the XFact dataset is highly imbalanced with 86% fake examples, the model achieves fairly good performance on the fake class, even in the claim-only setting, while performing poorly on the true class. Indeed, the model learns the underlying structure of claims in the dominant class (i.e., the fake class) more than those in the minority class. Therefore, providing the model with additional information about the context of each claim should be beneficial. Leveraging the evidence related to the claim significantly increases the classification performance for the true class, as shown in [14] and in our experimental results. Adding information about the credibility of the source provides the evidence further improves the classification performance, as shown in our experimental results.

Comparing with prior work, in [23], the authors combine representations of claim-related evidences and provide information about the sources in a one-hot encoding manner that would not measure the credibility of the sources. In [43], the authors consider a single credibility score for all claim-related articles based on their linguistic structure. By contrast, in our work we consider a vector of scores to evaluate the credibility of each evidence’s source. Our approach benefits from additional information provided by these scores since they are not obtained from the language structure of the evidence. Considering both claim-related evidences and additional information about the credibility of the source that produced the claim or the related evidence yields good performance and corroborates with previous studies. This highlights the importance of considering the context of the claim when building fake news detection models.

Furthermore, our approach yielded relatively good results on a Moroccan news dataset written in French and Arabic, that has not been used in the training phase. There are many interpretations and implications to these preliminary results. First, they demonstrate that the proposed approach can be used for the early detection of fake news in the early stages of propagation. Second, they suggest it is important to consider language-agnostic models when building a fake news detector. As shown in the ablation study, the evidence and its credibility scores contribute significantly to achieving such promising results on the Moroccan dataset.

As for the Constraint dataset, which is particularly curated for the classification of covid-19 related claims into fake and real, the proposed method achieves much better performance than the baseline and approaches the best result. Further, the performance obtained by our framework is remarkable given the fact the used Bert model is not fine-tuned on covid-19 related claims, unlike Bert ensemble used in [42]. The performance obtained by the entailment selection is lower than that obtained when considering all pieces of evidence, which suggests that the model benefits from richer context, and the evidence selection is an underlying process in the model. The performance degradation could also be due to the classification error induced by the entailment transformer, which may propagate in the upstream architecture.

In the light of the present results, enriching the feature space with the evidence and the source credibility score...
improves significantly the performance of fake claims detection. The combination of these elements with an emphasis on source credibility evaluation could be explored further in future studies to build robust models for early detection of fake news. In this sense, other source credibility metrics, besides the scores proposed here, could be further investigated in future studies. In fact, the source credibility could be used solely to validate or debunk a given claim.

V. CONCLUSION

In this paper, we present a multilingual framework for evaluating the veracity of online news based on available evidence and a credibility score assigned to the source. The evidence is found by conducting a Google search for the claim. We consider the top five search results as evidence articles and propose a measure to assess the credibility of the evidences’ sources based on SEO metrics including domain name authority and website reputation. The credibility metrics of each evidence were used as supplementary information that enhances the detection performance. To show the merits of our approach, we considered Covid-19 related news as a case study; we achieved an F1 score of 0.85 on the XFact dataset, outperforming existing methods. We also obtain an excellent detection result on the Constraint dataset with an F1 score of 0.97. Further, the proposed framework produces promising early detection results on a Moroccan news dataset that has not been used in the training phase. The method also has the advantage of being applicable in a multilingual setting, which allows it to be applied in a variety of contexts, regardless of the language. Besides, the study of claim-evidence entailment is also presented as an automatic evidence selection mechanism. This could be further investigated in future work in relation to the source credibility assessment.

REFERENCES

[1] X. Zhou and R. Zafarani, “A survey of fake news: Fundamental theories, detection methods, and opportunities,” ACM Comput. Surv., vol. 53, no. 5, pp. 1–40, Sep. 2020.
[2] A. Choudhary and A. Arora, “Linguistic feature based learning model for fake news detection and classification,” Expert Syst. Appl., vol. 169, May 2021, Art. no. 114171.
[3] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” ACM SIGKDD Explor. Newslett., vol. 19, no. 1, pp. 22–36, 2017.
[4] I. Singh, P. Deepak, and K. Anoop, “On the coherence of fake news articles,” in Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases, Cham, Switzerland: Springer, 2020, pp. 591–607.
[5] J. Pennebaker, R. Booth, R. Boyd, and M. Francis, Linguistic Inquiry and Word Count: LIWC2015, Austin, TX, USA: LIWC.Net, 2015.
[6] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, “Truth of varying shades: Analyzing language in fake news and political fact-checking,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2017, pp. 2931–2937.
[7] G. Pennycook and D. G. Rand, “Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking,” J. Personality, vol. 88, no. 2, pp. 185–200, Apr. 2020.
[8] G. Pennycook and D. G. Rand, “Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning,” Cognition, vol. 188, pp. 39–50, Jul. 2019.
[9] A. Ghanem, C. Assaad, H. Hafidi, Y. Moukafih, B. Guermah, N. Shibli, M. Zakroum, M. Ghogho, M. Dairi, M. Cherqaoui, and K. Baina, “Real-time infonveillance of Moroccan social media users’ sentiments towards the COVID-19 pandemic and its management,” Int. J. Environ. Res. Public Health, vol. 18, no. 22, Nov. 2021, Art. no. 12172.
[10] F. Haouari, M. Hasanain, R. Sawailah, and T. Elsayed, “ArCOV-19: The first Arabic COVID-19 Twitter dataset with propagation networks,” in Proc. 6th Arabic Natural Lang. Process. Workshop, 2021, pp. 82–91.
[11] F. Haouari, M. Hasanain, R. Sawailah, and T. Elsayed, “ArCOV-19-rumors: Arabic COVID-19 Twitter dataset for misinformation detection,” in Proc. 6th Arabic Natural Lang. Process. Workshop, 2021, pp. 72–81.
[12] W. S. Paka, R. Bansal, A. Kaushik, S. Sengupta, and T. Chakraborty, “Cross-SEAN: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection,” Appl. Soft Comput., vol. 107, Aug. 2021, Art. no. 107393.
[13] P. Patwa, S. Sharma, S. PYKL, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbai, A. Das, and T. Chakraborty, “Fighting an infodemic: COVID-19 fake news dataset,” in Proc. Int. Workshop Combating Lie Hostile Posts Regional Languages during Emergency Situation, 2020, pp. 21–29.
[14] A. Gupta and V. Srikanth, “X-fact: A new benchmark dataset for multilingual fact checking,” in Proc. 15th Int. Conf. Natural Lang. Process. (Short Papers), vol. 2, 2021, pp. 675–682.
[15] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, “SpotFake: A multi-modal framework for fake news detection,” in Proc. IEEE 5th Int. Conf. Multimedia Big Data (BigMM), Sep. 2019, pp. 39–47.
[16] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, “EANN: Explainable adversarial neural networks for multi-modal fake news detection,” in Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Jul. 2018, pp. 849–857.
[17] F. A. Ozbay and B. Alatas, “A novel approach for detection of fake news on social media using Metapheruristic optimization algorithms,” Elektronika Ir Elektrotechnika, vol. 25, no. 4, pp. 62–67, Aug. 2019.
[18] L. Cheng, R. Guo, K. Shu, and H. Liu, “Causal understanding of fake news dissemination on social media,” in Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining, Aug. 2021, pp. 148–157, doi: 10.1145/3447548.3467321.
[19] K. Shu, S. Wang, and H. Liu, “Beyond news contents: The role of social context for fake news detection,” in Proc. 12th ACM Int. Conf. Web Search Data Mining, Jan. 2019, pp. 312–320.
[20] X. Zhou and R. Zafarani, “Network-based fake news detection: A pattern-driven approach,” ACM SIGKDD Explor. Newsletter., vol. 21, no. 2, pp. 48–60, 2019.
[21] Y.-J. Lu and C.-T. Li, “GCAN: Graph-aware co-attention networks for explainable fake news detection on social media,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 505–514.
[22] K. Shu, D. Mahudeswaran, S. Wang, and H. Liu, “Hierarchical propagation networks for fake news detection: Investigation and exploitation,” in Proc. Int. AAAI Conf. Web Social Media, vol. 14, 2020, pp. 626–637.
[23] N. Vo and K. Lee, “Hierarchical multi-head attentive network for evidence-aware fake news detection,” in Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics, Main Volume, 2021, pp. 965–975.
[24] Y. Dou, K. Shu, C. Xia, P. S. Yu, and L. Sun, “User preference-aware fake news detection,” in Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Jul. 2021, pp. 2051–2055, doi: 10.1145/3404835.3462990.
[25] K. Shu, L. Cui, S. Xia, P. S. Yu, and L. Sun, “DEFEND: Explainable fake news detection,” in Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Jul. 2019, pp. 395–405.
[26] L. Hu, T. Yang, L. Zhang, W. Zhong, D. Tang, C. Shi, N. Duan, and M. Zhou, “Compare to the knowledge: Graph neural fake news detection with external knowledge,” in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process. (Long Papers), vol. 1, 2021, pp. 754–763.
[27] R. Vijjali, P. Potluri, S. Kumar, and S. Teki, “Two stage transformer model for COVID-19 fake news detection and fact checking,” in Proc. 3rd NLP4IF Workshop NLP Internet Freedom: Censorship, Disinformation, Propaganda, 2020, pp. 1–10.
[28] I. B. Schlicht and Z. Boukers, “ECOL: Early detection of COVID lies using content, prior knowledge and source information,” in Proc. CONRAIN@AAAI, 2021, pp. 141–152.
[29] Y. Wang, W. Yang, F. Ma, J. Xu, B. Zhong, Q. Deng, and J. Gao, “Weak supervision for fake news detection via reinforcement learning,” in Proc. AAAI Conf. Artif. Intell., vol. 34, 2020, pp. 516–523.
K. Popat, S. Mukherjee, A. Yates, and G. Weikum, “DeClarE: Debunking fake news and false claims using evidence-aware deep learning,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2018, pp. 22–32.