FbMultiLingMisinfo: Challenging Large-Scale Multilingual Benchmark for Misinformation Detection

Giorgio Barnabò
Department of Computer, Control and
Management Engineering Antonio Ruberti
Sapienza University of Rome
Rome, Italy
barnabo@diag.uniroma1.it

Federico Siciliano
Department of Computer, Control and
Management Engineering Antonio Ruberti
Sapienza University of Rome
Rome, Italy
siciliano@diag.uniroma1.it

Carlos Castillo
ICREA and
Pompeu Fabra University
Barcelona, Spain
chato@icrea.cat

Stefano Leonardi
Department of Computer, Control and
Management Engineering Antonio Ruberti
Sapienza University of Rome
Rome, Italy
leonardi@diag.uniroma1.it

Preslav Nakov
Qatar Computing Research Institute
HBKU
Doha, Qatar
pnakov@hbku.edu.qa

Giovanni Da San Martino
Dip. of Mathematics
University of Padova
Padua, Italy
dasan@math.unipd.it

Fabrizio Silvestri
Department of Computer, Control and
Management Engineering Antonio Ruberti
Sapienza University of Rome
Rome, Italy
fsilvestri@diag.uniroma1.it

2022 International Joint Conference on Neural Networks (IJCNN) | 978-1-7281-8671-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCNN55064.2022.9892739

***Abstract*—According to recent research, geometric deep learning allows to reach unprecedented accuracy for online misinformation detection. By fully leveraging the news social context, URL propagation paths in social networks are first represented as graphs and then classified using Graph Neural Network (GNN) models. Despite these remarkable efforts, researchers are still hampered by the scarcity of high-quality benchmark datasets, and as a result, the efficacy of state-of-the-art approaches could be overestimated. So far, in order to obtain a decent number of third-party fact-checked URLs, researchers have either sampled news from notoriously reliable and unreliable sources using distant supervision, or they have gathered pre-labeled URLs from third-party fact-checking websites. In the former case, resulting datasets can be quite large, but also noisy and biased since pieces of news are labeled as true or false according to their source label, and not individually fact-checked. In the latter case, assigned labels are more reliable, but the included news articles are usually in a single language and they may reflect unknown editorial decisions. As a result, datasets of the latter type are typically small, homogeneous, and thus unrealistically easy for automatic fake news detection models. In this work, we present FbMultiLingMisinfo, a new multilingual benchmark dataset, aimed at a more realistic evaluation of state-of-the-art misinformation detection models. URLs in our dataset come from the Facebook Privacy-Protected Full URLs Data Set, which we augmented with their propagation paths on Twitter. Our experimental results show that, when GNN-based models are tested on FbMultiLingMisinfo, recent misinformation detection results are only partially confirmed. We further show that a sharp reduction in the training size significantly reduces the model accuracy on FbMultiLingMisinfo, but not on two other widely used benchmark datasets for fake news detection.**

***Index Terms*—Misinformation, disinformation, fake news, fact-checking, factuality.**

I. Introduction

The most used and most cited benchmark datasets for misinformation detection come from fact-checking websites [1], [2], social network fact-checking units [3], or fake news threads in blogs [4], which researchers scrape in order to collect labeled news. When only a textual claim is given and the original URL is not explicitly provided, these fact-checked headlines are often automatically linked back to their likely original article, as in FakeNewsNet [1], procedure that can introduce further noise. Alternatively, a distant supervision strategy to collect larger corpora of semi-labeled news is to

This research was supported by the Italian Ministry of Education, University and Research (MIUR) under the grant “Dipartimenti di eccellenza 2018–2022” of the Department of Computer Science and the Department of Computer Engineering at Sapienza University of Rome. Partially supported by the ERC Advanced Grant 788893 AMDROMA “Algorithmic and Mechanism Design Research in Online Markets”, the EC H2020RIA project “SoBigData++” (871042), and the MIUR PRIN project ALGADIMAR “Algorithms, Games, and Digital Markets”. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Authorized licensed use limited to: Universita degli Studi di Roma La Sapienza. Downloaded on July 24,2023 at 12:51:50 UTC from IEEE Xplore. Restrictions apply.
978-1-7281-8671-9/22/$31.00 ©2022 IEEE

directly scrape articles from both trusted media outlets and fake news websites, without checking each piece of news individually [5]–[7]. Nonetheless, the assumption that news outlets only publish either real or fake news is very strong and not always realistic.

Moreover, such kind of distant supervision might yield highly correlated and homogeneous corpora. For instance, when articles are scraped from a list of news sources, it is usually fairly easy to predict their source, and, as a result, models often take a shortcut, rather than trying to solve the actual task [8]–[10]. Clearly, regardless of which of these two collecting procedures is chosen, resulting datasets do not represent a random sample of online information, and suffer from severely biased sampling strategies: news articles and URLs in standard datasets are mostly in English and are sampled during a specific period of time, over-represent fake news, and overall reflect unknown editorial decisions and inclusion criteria. This is why models trained on such benchmarks often do not generalize well to unseen news [11].

Over the past five years, research on misinformation detection has experienced significant momentum thanks to geometric deep learning. Graph Neural Networks (GNNs) enabled scientists to better model news diffusion patterns in social networks, thus moving away from simple text-based fake news detection pipelines. These state-of-the-art GNN-based misinformation detection methods try to classify graphs that represent URL cascades in social networks. Usually, given a piece of news and its corresponding URL, an attributed graph is created by connecting users and other entities that interacted and/or are related to the news. In most cases, involved entities are represented using rich node features, and finally Graph Neural Networks are applied on the resulting graphs, often yielding unprecedented levels of reported accuracy.

In this paper, we argue and empirically show that an intrinsic limitation of most research in misinformation detection, though, is the absence of large and high-quality benchmark datasets. Despite its decade-long history, current experiments are conducted on a few small and strongly biased datasets, which can easily lead to model efficacy overestimation. Data scarcity is mainly due to the high cost of third-party fact-checking, as well as to copyright issues. In order to determine whether a piece of news is false, an independent expert or a journalist has to analyse its content and has to provide written justification for their decision. Needless to say, the process is expensive and error-prone due to controversial topics and only partially false articles. Efforts to categorize different forms of misinformation and fake news are currently underway [12], but a clear and comprehensive framework is yet to emerge.

With the aim of helping researchers better assess misinformation detection methods, here we present *FbMultiLingMisinfo*, a new multilingual benchmark dataset that will enable a more realistic evaluation of state-of-the-art misinformation detection methods. We build our dataset starting from the recently published Facebook Privacy-Protected Full URLs Data Set [13], which comprises all 36 million URLs publicly shared on Facebook at least 100 times between January 2017 and July 2019, and includes third-party fact-checking labels for several of these URLs. As explained in Section III, in order to use this large set of multilingual pre-labeled URLs for misinformation detection purposes, we first had to gather their diffusion patterns. Since Facebook does not provide public APIs for research, we decided to leverage Twitter data instead. In particular, for each URL, we reconstructed its Twitter diffusion cascades and used them to test GNN-based misinformation detection models. Six state-of-the-art architectures were trained on our dataset, and results were compared to PolitiFact and GossipCop - two standard benchmark datasets released as part of FakeNewsNet [2]. To further analyze the differences between FbMultiLingMisinfo and existing benchmarks, we tested the performance of considered methods as the training size is gradually increased. Experiments show that our dataset is not only more difficult, but also more diverse as the decrease in training size has a much stronger impact compared to PolitiFact and GossipCop. To sum up, our key contributions are as follows:

- We present FbMultiLingMisinfo, a new challenging large-scale multilingual benchmark dataset for misinformation detection.
- We experiment with six state-of-the-art models, and we demonstrate that our dataset is more challenging than existing ones. In the best-case scenario, on FbMultiLingMisinfo we achieve 83% accuracy compared to 98% on GossipCop and 87% on PolitiFact.
- As for GossipCop, the largest fake news detection benchmark dataset, we further show that a very small training set allows to reach 97% accuracy with 5 models out of 6, putting into question its real discriminative power.
- Finally, we prove that GNN-based architectures do not always clearly outperform simpler sequential models based on Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs).

II. Related Work

In this section, we first review the most common benchmark datasets for misinformation detection, we then analyse different ways of representing news diffusion cascades through graphs, and finally, we focus on models for misinformation detection that leverage geometric deep learning.

A. Benchmark Datasets for Misinformation Detection

As mentioned before, in misinformation detection research, the validity of results strongly depends on the quality of the data used to conduct experiments. Unfortunately, fake news benchmark datasets are usually small and biased.

In most cases, only a few thousand third-party fact-checked URLs are used to test proposed methods, and these URLs are not randomly sampled from online information. A relatively large dataset coming from the fact-checking website gossipcop.com, and a smaller one sampled from politifact.com - both released as part of the FakeNewsNet [1] - constitute two of the most commonly used benchmark datasets [1], [11], [14]–[20]. Other common sources of annotated URLs or posts are

Authorized licensed use limited to: Universita degli Studi di Roma La Sapienza. Downloaded on July 24,2023 at 12:51:50 UTC from IEEE Xplore. Restrictions apply.

BuzzFeed [14], [16], [17], Twitter [21]–[23] and Weibo [23], [24]. In these datasets, assigned labels are usually reliable, but included news are in a single language and reflect unknown selection and factuality criteria, which means that resulting datasets are typically small, homogeneous, and easy to beat. As an alternative strategy to obtain a larger number of third-party fact-checked URLs, researchers have also sampled news from notoriously reliable and unreliable sources using distant supervision. Resulting datasets, such as NELA [5], [6] can be quite large, but they are also noisy and biased since news articles are labeled as true or false according to their source label, and are not individually fact-checked. To conclude, it is worth mentioning that an extensive and in-depth review of existing evaluation datasets was recently published [25].

B. News Cascades Representation

Concerning the entities used as nodes, in most cases news articles themselves represent a natural choice [14]–[16], [20], [26], [27], followed by users who published a specific URL/post in the social network [11], [16], [20]–[22], [26]–[29]. Since currently only Twitter gives easy access to its API for research purposes, almost all papers look at Twitter users posting a news URL, or re-posting a specific tweet [22], [28], [30]. Together with news, posts, and users, content creators represent another important signal and are often included as separate nodes [14]–[16], [20], [26], [29]. Moreover, when available, the article or the post authors can also be used; otherwise, the source represents a valid alternative [20], [26]. Finally, some less popular choices are topics [14], [15] and post comments [21]. As for the edges, news articles can be directly connected to their creators [14]–[16], [20], [26], to their topic [14], [15], and to their posting users [26]. Users, in turn, can be linked through their social graph, e.g., based on following or friendship relationships [20], [26], [28], or re-posting activity [27]. The cosine similarity between a vector representation for two users is another possible option [27]. Moreover, users can be connected to their posts [16], [21], to an article through a stance score [20], [26], and to nodes representing their posted comments [21], which in turn are usually connected to their corresponding post [21]. As for news-posting domains, an edge can be added every time two posting domains link to each other [20], [26]. Finally, when the tweets are nodes themselves, an edge could mean that tweet i is responding to tweet j, like in [30].

C. GNN-based Misinformation Detection Methods

As extensively shown in a seminal work [31], in social networks fake news spreads faster and deeper than high-quality information. This new awareness rapidly lead to a major shift in how scientists tackle the difficult task of misinformation detection. Indeed, over the past five years, GNN-based methods have established themselves as the new state-of-the-art approaches in the fight against fake news. Unlike their content-based predecessors, these methods leverage the diffusion patterns of news in social networks as their main discriminant signal. Initially, as an intermediate step, features representing the spreading patterns of the news were simply added to content-based features in order to train traditional machine learning classifiers. For example, researchers started using Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) to classify the sequence of users posting some specific news [23]. More recently, the task has been typically formulated as a node [14], [20]–[22], [26], [27] or a graph classification task [11], [28], [30]. That is, existing methods make use of either node or graph embeddings obtained by training a geometric deep learning architecture on customized graphs. The most commonly used architectures include Graph Convolution Networks (GCN) [22], [28], [32], Bi-Directional Graph Convolution Networks (BiGCN) [33], Graph Attention Networks [14], [32], [34], [35], and GraphSAGE [11], [32]. Depending on the approach, these latent representations can be further combined with other text-based features and/or with non-GNN-based embeddings that capture other aspects of fake news diffusion [22].

III. FbMultiLingMisinfo & Comparison Datasets

Here we present FbMultiLingMisinfo, a new multilingual collection of third-party fact-checked news and their diffusion cascades on Twitter. To create the dataset, we started from the Facebook Privacy-Protected Full URLs Data Set [13], which was recently published and includes any URL publicly shared on Facebook at least 100 times between January 2017 and July 2019. About 12,500 URLs out of a total of 36 millions URLs come with a third-party fact-checking label, and we expanded this subset with additional information from Twitter. Our goal is to provide researchers with a new challenging playground to test automatic misinformation detection models. Although the original Facebook dataset was used in previous work [36]–[40], to the best of our knowledge we are the first to use it for misinformation detection. While still preserving some shortcomings of all misinformation detection datasets, FbMultiLingMisinfo introduces several improvements with respect to current benchmarks:

- it is the first multilingual dataset for misinformation detection;
- it is the second largest benchmark dataset for misinformation detection whose URLs were individually third-party fact-checked;
- it is more complex than PolitiFact and GossipCop, the two most used benchmark datasets for misinformation detection;
- all included URLs are highly impactful (shared at least 100 times on Facebook);
- compared to GossipCop and PolitiFact, many more Twitter users were involved in spreading of the included URLs as shown in section III.

In order to test state-of-the-art misinformation detection methods on our new dataset, the first thing we had to do was to reconstruct the diffusion graphs of the included URLs. Thanks to the Twitter API for Academic Research, we were able to download all Twitter posts containing each of the 12,447 fact-checked URLs, as well as several features about all the authors

Authorized licensed use limited to: Universita degli Studi di Roma La Sapienza. Downloaded on July 24,2023 at 12:51:50 UTC from IEEE Xplore. Restrictions apply.

TABLE I
STATISTICS ABOUT THE FIVE DATASETS: FBMULTILINGMISINFO, GOSSIPCOP LARGE, GOSSIPCOP SMALL, POLITIFACT LARGE, POLITIFACT SMALL

Dataset	Fake News	True News	Total News	Twitter Posts	Twitter Users	# Posts / # Users
FbMulti LingMisinfo	4,034	3,300	7,334	3,219,383	1,240,592	2,59
GossipCop Large	3,942	13,577	17,519	1,272,256	281,962	4,50
GossipCop Small	2,732	2,732	5,464	236,889	31,101	7,61
PolitiFact Large	293	305	598	495,202	293,626	1,69
PolitiFact Small	157	157	314	22,340	14,873	1,50

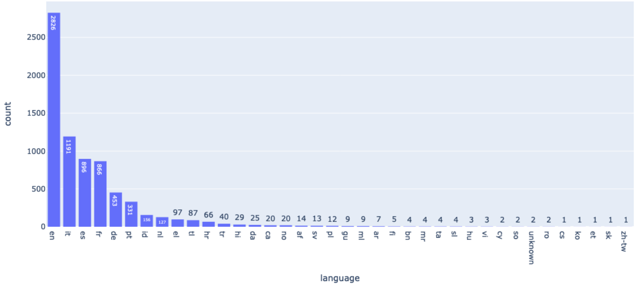

Fig. 1. FbMultiLingMisinfo: news articles language distribution (ISO 639-1 codes).

of these posts. Unfortunately, Twitter does not give access to deleted posts and does not allow to retrieve statistics about a post's author at the time a URL was posted. In other words, collected data represent *current* statistics and information about users who posted certain URLs *in the past*. This is an intrinsic limitation of virtually all previously published research leveraging Twitter data for fake news detection, and as such it is difficult to overcome.

Finally, in order to make our analysis more sound and to focus on high-impact news, we decided to apply several pre-processing steps to our raw dataset. First, we detected and deleted duplicates using multilingual sentence embeddings [41]. Namely, we considered titles with a cosine similarity greater than 0.9 as duplicates (see Section III-B for more details). In case of duplicates, we kept the earliest URL reporting a news, and we discarded all duplicates that came afterwards. The rationale is that once an automatic system has discovered that a given piece of news is fake news, it would be relatively easy to discard similar news later, e.g. based on textual similarity [42], [43]. Moreover, we only included URLs with at least four posts on Twitter, and we deleted URLs that point to fact-checking websites. Statistics about the final datasets are shown in Table I, while Fig. 1 shows the distribution of languages in the final list of 7,334 URLs, and Table II lists the top-10 most frequent domains. As we will show below, our dataset represents a new hard-to-beat benchmark for current misinformation detection approaches.

TABLE II
FBMULTILINGMISINFO: TOP-10 DOMAINS AND THE CORRESPONDING CUMULATIVE PERCENTAGE OF URLS THEY COVER.

Domain	Fact-Checked URLs	Cumulative Percentage
www.repubblica.it	151	2%
www.youtube.com	106	4%
www.ilfattoquotidiano.it	106	5%
www.breitbart.com	72	6%
www.ansa.it	66	7%
yournewswire.com	59	8%
www.corriere.it	56	8%
video.repubblica.it	50	9%
www.huffingtonpost.it	48	10%
www.pulzo.com	47	10%

TABLE III
GOSSIPCOP: TOP-10 DOMAINS AND THE CORRESPONDING CUMULATIVE PERCENTAGE OF URLS THEY COVER.

Domain	Fact-Checked URLs	Cumulative Percentage
people.com	1,403	8%
www.dailymail.co.uk	737	12%
www.usmagazine.com	550	15%
www.etonline.com	544	18%
www.longroom.com	465	21%
en.wikipedia.org	459	24%
hollywoodlife.com	435	26%
www.hollywoodreporter.com	269	28%
www.usatoday.com	242	29%
variety.com	237	30%

In our GitHub repository,[1], for each URL included in FbMultiLingMisinfo, we listed all the tweet IDs and Twitter user IDs making up our dataset. In order to replicate our experiments and/or to use FbMultiLingMisinfo, researchers must first apply to Social Science One[2] and obtain the Facebook Privacy-Protected Full URLs Data Set. Then, they have to filter all the URLs included in FbMultiLingMisinfo with the provided URL IDs, and finally they can hydrate tweet and Twitter user IDs through the Twitter APIs.

TABLE IV
POLITIFACT: TOP-10 DOMAINS AND THE CORRESPONDING CUMULATIVE PERCENTAGE OF URLS THEY COVER.

Domain	Fact-Checked URLs	Cumulative Percentage
web.archive.org	94	16%
www.youtube.com	22	19%
www.politifact.com	17	22%
www.washingtonpost.com	14	25%
www.nytimes.com	13	27%
yournewswire.com	10	28%
www.cnn.com	9	30%
www.cq.com	8	31%
www.whitehouse.gov	7	32%
abcnews.go.com	7	34%

Authorized licensed use limited to: Universita degli Studi di Roma La Sapienza. Downloaded on July 24,2023 at 12:51:50 UTC from IEEE Xplore. Restrictions apply.

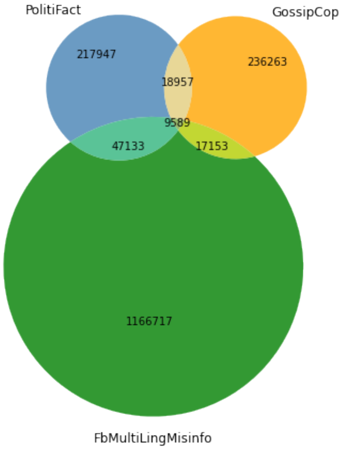

Fig. 2. Overlapping Twitter users spreading the URLs in each of the three datasets: FbMultiLingMisinfo, GossipCop-Large, and PolitiFact-Large.

A. Comparison Datasets

In order to show that FbMultiLingMisinfo is a more challenging benchmark than pre-existing datasets, we tested several state-of-the-art misinformation detection methods on it, and we showed that they perform much worse than on PolitiFact and GossipCop, the two most widely used benchmarks for fake news detection. PolitiFact and GossipCop come from FakeNewsNet [1] and were collected from two fact-checking websites that focus on celebrities and political reporting, respectively. To make results comparable, before running the experiments, we applied the same pre-processing steps listed for FbMultiLingMisinfo.

Tables III and IV show that GossipCop and PolitiFact are much more homogeneous in terms of publishing domains compared to FbMultiLingMisinfo, with 30% of their URLs coming from just ten domains, while this figure drops to 10% for FbMultiLingMisinfo. Moreover, as shown in Table I and Fig. 2, URLs included in our FbMultiLingMisinfo were spread by many more Twitter users compared to those in GossipCop and PolitiFact. In other words, while GossipCop-Large is more than twice as large as FbMultiLingMisinfo in terms of number of news articles, its URLs were spread by five times less unique Twitter users, thus making the classification task particularly easy. Finally, Fig.3 and Fig.4 show differences in the temporal distribution and impact between the three datasets. On average, the URLs in FbMultiLingMisinfo spread deeper and were first published between 2017 and 2019, while GossipCop and PolitiFact stopped sampling new URLs at the end of 2018.

For data availability reasons, we used two different versions of GossipCop and PolitiFact, depending on the tested architecture. As further explained in the next section, for the only non-GNN-based model, we ran experiments on GossipCop-Large and PolitiFact-Large, while for GNN-based models, we had to rely on GossipCop-Small and PolitiFact-Small, which are subset of GossipCop-Large and PolitiFact-Large (statistics shown in Table I). We could not run the GNN-based models on GossipCop-Large and PolitiFact-Large for two reasons: (*i*) much of the Twitter data was not available anymore, (*ii*) reconstructing the diffusion cascades as done in [32] has proven quite hard. Thus, we relied on the diffusion graphs shared by [32].

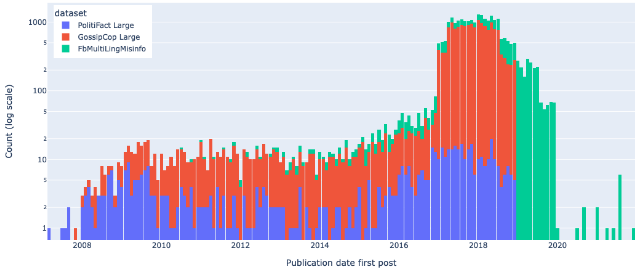

Fig. 3. Temporal distribution of URLs for FbMultiLingMisinfo, GossipCop-Large, and PolitiFact-Large.

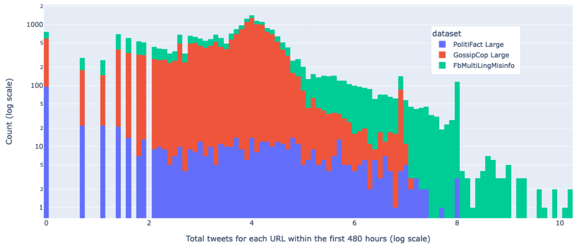

Fig. 4. Histogram of the total number of tweets for each URL within the first 480 hours.

As for FbMultiLingMisinfo, information about URLs used in our experiments and corresponding Twitter data can be found in our GitHub repository,[3] while researchers can go to the FakeNewsNet GitHub repository[4] to obtain the full versions of both GossipCop and PolitiFact.

B. Modeling the Diffusion Cascades of URLs Shared on Twitter

The models we experimented with take as input two different kinds of data. In one case [23], the sequence of Twitter users who posted a URL is fed to an RNN-based model, which then predicts whether the corresponding news is fake or real. In order to use this model, we first downloaded the full list of tweets that shared each URL, and we then represented each piece of news as a sequence of corresponding Twitter users. Similarly to [23], [32], each Twitter user is represented as a vector of 13 features: length of the user name, user description and username, number of followers and followings, total number of tweets and lists, time passed since

[1] http://github.com/siciliano-diag/FbMultiLingMisinfo

[2] http://socialscience.one/

[3] http://github.com/siciliano-diag/FbMultiLingMisinfo

[4] http://github.com/KaiDMML/FakeNewsNet

Authorized licensed use limited to: Universita degli Studi di Roma La Sapienza. Downloaded on July 24,2023 at 12:51:50 UTC from IEEE Xplore. Restrictions apply.

the account creation, and finally whether the account has a URL, a location, a profile image, and whether it is protected and/or verified.

In contrast, for GNN-based models, we had to construct a graph representing each URL diffusion cascades. As in [32], given the sequence of tweets and retweets mentioning a URL, we built a graph as follows: a central node represents the news and there is an additional node for each tweet. All direct tweets are connected to the central node, while re-tweets are connected to the tweet they are re-tweeting. Finally, similarly to [32], we obtained the node features by encoding the user description with the paraphrase-multilingual-mpnet-base-v2 model from the Hugging Face multilingual sentence embedding model trained as in [41].

For the central node representing the URL, we used the news title embedding. Our choice of a multilingual model is due to the multilingual nature of the FbMultiLingMisinfo dataset. For the GossipCop and PolitiFact datasets, we used the diffusion graphs shared in [32], but we replaced the given node features with the multilingual sentence embeddings.

IV. Experiments & Results

In this section, we describe the experiments we conducted using several state-of-the-art misinformation detection models on our dataset, as well as on GossipCop and PolitiFact.

A. Experiments

We experimented with six state-of-the-art approaches, which use news contextual and social information in two different ways. A non-GNN-based method that leverages the sequence of Twitter users posting a URL, and five GNN-based methods that work on news diffusion graphs.

- CNN-RNN [23] This model integrates an RNN-based section, using a Gated Recurrent Unit (GRU), and a CNN-based section applied to the user features. In our experiments, we considered the first 100 tweets for each URL.
- GCNFN [28] This is the first model that applied graph convolution networks on news diffusion graphs. The model also makes use of more recent developments in the field of neural networks, such as graph attention for dimensionality reduction, Scaled Exponential Linear Unit (SELU) as non-linearity and Hinge loss.
- BiGCN [33] A newer and more advanced bidirectional version of Graph Convolutional Networks (GCN) that aims at capturing news diffusion and news dispersion simultaneously.
- GCN [44] A simple GCN that uses an efficient layer-wise propagation rule based on a first-order approximation of spectral convolutions on graphs. It can learn hidden layer representations that encode both local graph structure and features of nodes.
- GAT [45] The use of multi-head graph attention makes this model computationally highly efficient, thus allowing it to deal with neighbourhoods of various sizes without depending on knowing the entire graph structure upfront.

TABLE V
Summary of our experiments.

Model	Datasets	Train Sizes	Work Using the Model for Misinformation Detection
CNN-RNN	FbMultiLingMisinfo GossipCop Small PolitiFact Small	5%, 10%, 20%, 50%	[23]
GCNFN	FbMultiLingMisinfo GossipCop Large PolitiFact Large		[28], [32]
BiGCN			[32], [33]
GCN			[22], [32]
GAT			[14], [29], [32], [34]
GraphSAGE			[11], [32]

- GraphSAGE [46] This model exploits inductive node embedding by making use of node features in order to generalise to unseen nodes.

Implementation-wise, for the CNN-RNN model, we used the code from [23],[5] while for GNN-based architectures, we used the code distributed by [32].[6] Concerning the hyperparameters, we used the values from the original papers.

We tested the six models on FbMultiLingMisinfo, GossipCop and PolitiFact with increasing training sizes: 5%, 10%, 20%, 50% of the training data. We repeated each experiment five times with different random seeds and we report averaged results. Table V shows a summary of our experiments.

B. Results

Starting from a small training size allowed us to see some important differences between the three datasets.

- First, as shown in Table VI, regardless of the training size or model chosen, the accuracy on GossipCop was always above 95%, except for BiGCN. This shows that GossipCop does not allow to discriminate well between different models, as 5% of the training data is enough to achieve an accuracy higher than 97%. A likely explanation is that a relatively small number of unique Twitter users spread all the URLs included in GossipCop as shown in Table I. Unlike FbMultiLingMisinfo and PolitiFact, whose average number of post per user is 2.59 and 1.50 respectively, for GossipCop-Large and GossipCop-Small the same figure rises to 4.5 and 7.6.
- On the other hand, for both PolitiFact and FbMultiLingMisinfo, increasing the training size always yielded noticeable gains in performance for all models, going from around 74% up to 87% for PolitiFact and from around 74% up to 82% for FbMultiLingMisinfo (Table VI). Fig. 5 further shows the best model accuracy for each training size on FbMultiLingMisinfo, GossipCop and PolitiFact. FbMultiLingMisinfo is clearly harder than PolitiFact and GossipCop, and needs more training data to reach somewhat good (but not astonishing) accuracy.
- Indeed, even when 50% of the URLs are used for training, accuracy on FbMultiLingMisinfo does not exceed 83%, which puts into question the real validity of the proposed

[5] http://github.com/yumere/early-fakenews-detection

[6] http://github.com/safe-graph/GNN-FakeNews

Authorized licensed use limited to: Universita degli Studi di Roma La Sapienza. Downloaded on July 24,2023 at 12:51:50 UTC from IEEE Xplore. Restrictions apply.

TABLE VI

EVALUATION RESULTS WHEN USING 5%, 10%, 20%, AND 50% OF THE TRAINING DATA. THE TEST SET IS KEPT CONSTANT AND REPRESENT 30% OF THE DATASET. THE REMAINING URLS ARE USED AS A VALIDATION SET.

Model	**PolitiFact**		**GossipCop**		**FbMulti LingMisinfo**	
	Acc.	F1	Acc.	F1	Acc.	F1
Train Set Size: 5%						
CNN-RNN	0.74	0.74	0.95	0.88	0.73	0.77
GCNFN	0.73	0.70	0.95	0.95	0.75	0.78
BiGCN	0.74	0.75	0.81	0.81	0.75	0.78
GCN	0.72	0.79	0.96	0.96	0.74	0.77
GAT	0.67	0.63	0.95	0.95	0.70	0.73
GraphSAGE	0.78	0.78	0.97	0.97	0.74	0.76
Train Set Size: 10%						
CNN-RNN	0.77	0.75	0.96	0.90	0.75	0.78
GCNFN	0.76	0.77	0.96	0.96	0.78	0.79
BiGCN	0.82	0.79	0.80	0.81	0.78	0.80
GCN	0.78	0.77	0.96	0.96	0.76	0.78
GAT	0.81	0.80	0.97	0.97	0.74	0.77
GraphSAGE	0.75	0.75	0.97	0.97	0.76	0.79
Train Set Size: 20%						
CNN-RNN	0.77	0.75	0.96	0.92	0.76	0.79
GCNFN	0.73	0.71	0.96	0.96	0.78	0.81
BiGCN	0.84	0.84	0.89	0.89	0.79	0.82
GCN	0.79	0.79	0.97	0.97	0.79	0.81
GAT	0.82	0.81	0.97	0.97	0.79	0.81
GraphSAGE	0.79	0.79	0.97	0.97	0.79	0.81
Train Set Size: 50%						
CNN-RNN	0.81	0.80	0.97	0.94	0.77	0.80
GCNFN	0.87	0.87	0.97	0.97	0.82	0.83
BiGCN	0.84	0.86	0.95	0.95	0.82	0.84
GCN	0.79	0.78	0.98	0.98	0.82	0.84
GAT	0.84	0.84	0.98	0.98	0.83	0.85
GraphSAGE	0.80	0.75	0.98	0.98	0.83	0.85

methods. Even though accuracy over 80% could still be considered a good result, implementing such methods in a real social network could result in censoring a large number of real news.

- Another very interesting observation drawn from Table VI is that, for all training sizes below 50% (5%, 10% and 20%), none of the GNN-based models significantly outperform the RNN-CNN model. Only on PolitiFact, some GNN-based models perform better than RNN+CNN, but the difference in performance never exceeds 4%. In other words, an intrinsically sequential architecture seems to capture as much information as more advanced geometric deep learning methods, thus questioning their true hegemony in this domain.

V. CONCLUSIONS & FUTURE WORK

We presented FbMultiLingMisinfo, a new large-scale multilingual benchmark dataset for misinformation detection, and we tested several state-of-the-art GNN-based models on it. The results show that our new dataset is more challenging than GossipCop and PolitiFact, two widely used fake news detection benchmarks from FakeNewsNet [1]. GossipCop, in particular, seems to be exceptionally easy to classify and thus of limited utility to assess the discriminatory power of misinformation detection methods; PolitiFact is slightly harder, but also much smaller. Testing the models with different and increasing training sizes allowed us to further differentiate the three datasets and different architectures. A noteworthy result is that an intrinsically sequential model based on RNN and CNN performs very similarly to state-of-the-art geometric deep learning architectures when up to 20% of the data is used for training.

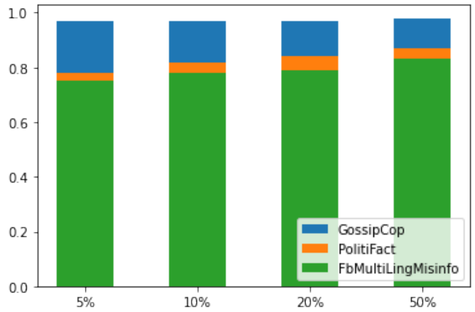

Fig. 5. Best model accuracy for FbMultiLingMisinfo, GossipCop and PolitiFact, for each training size. We can see that for GossipCop increasing the training size does not affect much accuracy, which is already above 97% when using just 5% of the training data.

In future work, as the Facebook Privacy-Protected Full URLs Data Set is constantly updated, we plan to keep enriching FbMultiLingMisinfo with newly released URLs, thus creating a constantly growing high-quality benchmark dataset for misinformation detection. In addition, in order to test a more realistic setting, differences between datasets and tested architectures could be assessed for early detection, and methods could be probed with highly unbalanced datasets or under fact-checking budget constraints. In the latter case, both offline and online active learning strategies could come in handy and would help define an optimal strategy to continuously fine-tune the most successful methods [14], [47]. Finally, another interesting direction could be to build ensembles aggregating results from a variety of misinformation detection models.

REFERENCES

[1] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big data*, vol. 8, no. 3, pp. 171–188, 2020.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[3] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," 2016.

[4] K. Nakamura, S. Levy, and W. Y. Wang, "r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," *arXiv preprint arXiv:1911.03854*, 2019.

[5] J. Nørregaard, B. D. Horne, and S. Adalı, "Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles," in *Proceedings of the international AAAI conference on web and social media*, vol. 13, 2019, pp. 630–638.

[6] M. Gruppi, B. D. Horne, and S. Adalı, "Nela-gt-2020: A large multi-labelled news dataset for the study of misinformation in news articles," *arXiv preprint arXiv:2102.04567*, 2021.

[7] L. Bozarth and C. Budak, "Toward a better performance evaluation framework for fake news classification," in *Proceedings of the international AAAI conference on web and social media*, vol. 14, 2020, pp. 60–71.

Authorized licensed use limited to: Universita degli Studi di Roma La Sapienza. Downloaded on July 24,2023 at 12:51:50 UTC from IEEE Xplore. Restrictions apply.

[8] A. Barrón-Cedeno, I. Jaradat, G. Da San Martino, and P. Nakov, “Proppy: Organizing the news based on their propagandistic content,” *Information Processing & Management*, vol. 56, no. 5, pp. 1849–1864, 2019.

[9] J. Kiesel, M. Mestre, R. Shukla, E. Vincent, P. Adineh, D. Corney, B. Stein, and M. Potthast, “SemEval-2019 task 4: Hyperpartisan news detection,” in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA, Jun. 2019, pp. 829–839.

[10] R. Baly, G. Da San Martino, J. Glass, and P. Nakov, “We can detect your bias: Predicting the political ideology of news articles,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Nov. 2020, pp. 4982–4991.

[11] Y. Han, S. Karunasekera, and C. Leckie, “Graph neural networks with continual learning for fake news detection from social media,” *arXiv preprint arXiv:2007.03316*, 2020.

[12] M. R. Islam, S. Liu, X. Wang, and G. Xu, “Deep learning for misinformation detection on online social networks: a survey and new perspectives,” *Social Network Analysis and Mining*, vol. 10, no. 1, pp. 1–20, 2020.

[13] S. Messing, C. DeGregorio, B. Hillenbrand, G. King, S. Mahanti, Z. Mukerjee, C. Nayak, N. Persily, B. State, and A. Wilkins, “Facebook Privacy-Protected Full URLs Data Set,” 2020. [Online]. Available: https://doi.org/10.7910/DVN/TDOAPG

[14] Y. Ren, B. Wang, J. Zhang, and Y. Chang, “Adversarial active learning based heterogeneous graph neural network for fake news detection,” in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 452–461.

[15] J. Zhang, B. Dong, and S. Y. Philip, “Fakedetector: Effective fake news detection with deep diffusive neural network,” in *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 2020, pp. 1826–1829.

[16] K. Shu, S. Wang, and H. Liu, “Beyond news contents: The role of social context for fake news detection,” in *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019, pp. 312–320.

[17] X. Zhou and R. Zafarani, “Network-based fake news detection: A pattern-driven approach,” *ACM SIGKDD explorations newsletter*, vol. 21, no. 2, pp. 48–60, 2019.

[18] K. Shu, D. Mahudeswaran, S. Wang, and H. Liu, “Hierarchical propagation networks for fake news detection: Investigation and exploitation,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 626–637.

[19] K. Shu, G. Zheng, Y. Li, S. Mukherjee, A. H. Awadallah, S. Ruston, and H. Liu, “Leveraging multi-source weak social supervision for early detection of fake news,” *arXiv preprint arXiv:2004.01732*, 2020.

[20] S. Chandra, P. Mishra, H. Yannakoudakis, M. Nimishakavi, M. Saeidi, and E. Shutova, “Graph-based modeling of online communities for fake news detection,” *arXiv preprint arXiv:2008.06274*, 2020.

[21] J. Yu, Q. Huang, X. Zhou, and Y. Sha, “Iarnet: An information aggregating and reasoning network over heterogeneous graph for fake news detection,” in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–9.

[22] Y.-J. Lu and C.-T. Li, “Gcan: Graph-aware co-attention networks for explainable fake news detection on social media,” *arXiv preprint arXiv:2004.11648*, 2020.

[23] Y. Liu and Y.-F. B. Wu, “Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks,” in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[24] A. Lao, C. Shi, and Y. Yang, “Rumor detection with field of linear and non-linear propagation,” in *Proceedings of the Web Conference 2021*, 2021, pp. 3178–3187.

[25] A. D’Ulizia, M. C. Caschera, F. Ferri, and P. Grifoni, “Fake news detection: a survey of evaluation datasets,” *PeerJ Computer Science*, vol. 7, p. e518, 2021.

[26] V.-H. Nguyen, K. Sugiyama, P. Nakov, and M.-Y. Kan, “Fang: Leveraging social context for fake news detection using graph representation,” in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1165–1174.

[27] A. Silva, Y. Han, L. Luo, S. Karunasekera, and C. Leckie, “Propagation2vec: Embedding partial propagation networks for explainable fake news early detection,” *Information Processing & Management*, vol. 58, no. 5, p. 102618, 2021.

[28] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, “Fake news detection on social media using geometric deep learning,” *arXiv preprint arXiv:1902.06673*, 2019.

[29] C. Yuan, Q. Ma, W. Zhou, J. Han, and S. Hu, “Early detection of fake news by utilizing the credibility of news, publishers, and users based on weakly supervised learning,” *arXiv preprint arXiv:2012.04233*, 2020.

[30] C. Song, K. Shu, and B. Wu, “Temporally evolving graph neural network for fake news detection,” *Information Processing & Management*, vol. 58, no. 6, p. 102712, 2021.

[31] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[32] Y. Dou, K. Shu, C. Xia, P. S. Yu, and L. Sun, *User Preference-Aware Fake News Detection*. New York, NY, USA: Association for Computing Machinery, 2021, p. 2051–2055. [Online]. Available: https://doi.org/10.1145/3404835.3462990

[33] T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, and J. Huang, “Rumor detection on social media with bi-directional graph convolutional networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 549–556.

[34] Q. Huang, J. Yu, J. Wu, and B. Wang, “Heterogeneous graph attention networks for early detection of rumors on twitter,” in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[35] Y. Ren and J. Zhang, “Fake news detection on news-oriented heterogeneous information networks through hierarchical graph attention,” in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[36] A. Bailey, T. Gregersen, and F. Roesner, “Interactions with potential mis/disinformation urls among us users on facebook, 2017-2019,” in *Proceedings of the ACM SIGCOMM 2021 Workshop on Free and Open Communications on the Internet*, 2021, pp. 16–26.

[37] C. Buntain, R. Bonneau, J. Nagler, and J. A. Tucker, “A summary statistic for political ideology of web domains using differentially private engagement data with applications to social science one,” *Available at SSRN 3765240*, 2021.

[38] P. Ortellado, M. Moretto Ribeiro, G. Kessler, G. Vommaro, J. C. Rodriguez-Raga, J. P. Luna, E. Heinen, L. F. Cely, and S. Toro, “Old adults are more engaged on facebook, especially in politics: Evidence from users in 46 countries,” *Especially in Politics: Evidence From Users in*, vol. 46, 2021.

[39] J. Allen, M. Mobius, D. M. Rothschild, and D. J. Watts, “Research note: Examining potential bias in large-scale censored data,” *Harvard Kennedy School Misinformation Review*, 2021.

[40] G. Evans and G. King, “Statistically valid inferences from differentially private data releases, with application to the facebook urls dataset,” *Political Analysis. URL: GaryKing. org/dpd*, 2020.

[41] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[42] S. Shaar, N. Babulkov, G. Da San Martino, and P. Nakov, “That is a known lie: Detecting previously fact-checked claims,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 3607–3618.

[43] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, and T. Mandl, “The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news,” in *Advances in Information Retrieval*, D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, and F. Sebastiani, Eds. Cham: Springer International Publishing, 2021, pp. 639–649.

[44] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.

[45] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, “Graph attention networks,” *arXiv preprint arXiv:1710.10903*, 2017.

[46] W. L. Hamilton, R. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1025–1035.

[47] P. Ksieniewicz, P. Zyblewski, M. Choraś, R. Kozik, A. Giełczyk, and M. Woźniak, “Fake news detection from data streams,” in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

Authorized licensed use limited to: Universita degli Studi di Roma La Sapienza. Downloaded on July 24,2023 at 12:51:50 UTC from IEEE Xplore. Restrictions apply.