Information Sciences 615 (2022) 657–677

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

In-context annotation of topic-oriented datasets of fake news: A case study on the notre-dame fire event

Lucia C. Passaro [a,*], Alessandro Bondielli [a,c], Pietro Dell'Oglio [b,d], Alessandro Lenci [c], Francesco Marcelloni [b]

[a] Department of Computer Science, University of Pisa, Largo B. Pontecorvo, 3, Pisa, Italy
[b] Department of Information Engineering, University of Pisa, Largo Lucio Lazzarino, 1, Pisa, Italy
[c] Department of Philology, Literature and Linguistics, University of Pisa, Piazza Torricelli, 2, Pisa, Italy
[d] Department of Information Engineering, University of Florence, Via di S. Marta, 3, Florence, Italy

ARTICLE INFO

Article history:
Received 4 June 2021
Received in revised form 26 October 2021
Accepted 23 July 2022
Available online 28 July 2022

Keywords:
Fake news
Dataset collection and annotation
Machine learning

ABSTRACT

The problem of fake news detection is becoming increasingly interesting for several research fields. Different approaches have been proposed, based on either the content of the news itself or the context and properties of its spread over time, specifically on social media. In the literature, it does not exist a widely accepted general-purpose dataset for fake news detection, due to the complexity of the task and the increasing ability to produce fake news appearing credible in particular moments. In this paper, we propose a methodology to collect and label news pertinent to specific topics and subjects. Our methodology focuses on collecting data from social media about real-world events which are known to trigger fake news. We propose a labelling method based on crowdsourcing that is fast, reliable, and able to approximate expert human annotation. The proposed method exploits both the content of the data (i.e., the texts) and contextual information about fake news for a particular real-world event. The methodology is applied to collect and annotate the Notre-Dame Fire Dataset and to annotate part of the PHEME dataset. Evaluation is performed with fake news classifiers based on Transformers and fine-tuning. Results show that context-based annotation outperforms traditional crowdsourcing out-of-context annotation.

1. Introduction

Given the social and political impact of disinformation and misinformation, *fake news detection* has become a widely researched topic. The interest in fake news and *rumours* has grown considerably in the last few years [1]. The problem has been analyzed from several different perspectives, including *Natural Language Processing* (NLP), *Data Mining* (DM) and *Social Network Analysis* (SNA).

One of the most widely accepted definitions of fake news, given by [2], refers to "a news article that is intentionally and verifiably false". On the other hand, rumours are often referred to as "circulating stories of questionable veracity, which are apparently credible but hard to verify, and produce sufficient skepticism and/or anxiety" [3]. We can argue that this

* Corresponding author.
E-mail addresses: lucia.passaro@unipi.it (L.C. Passaro), alessandro.bondielli@unipi.it (A. Bondielli), pietro.delloglio@unifi.it (P. Dell'Oglio), alessandro.lenci@unipi.it (A. Lenci), francesco.marcelloni@unipi.it (F. Marcelloni).

https://doi.org/10.1016/j.ins.2022.07.128
0020-0255/

definition is relevant for fake news as well, especially in the context of social media. Pieces of misinformation or disinformation that we can generally refer to as fake news are in fact mostly spread through social media platforms such as Facebook and Twitter, often in the form of posts linking to fake articles or decontextualized news used deceitfully. This is especially true in the case of real-world events. Such events are in fact a fertile ground for the spread of false or unverified information. For example, during the catastrophic fire that struck the Notre-Dame De Paris Cathedral in April 2019, several fake news were spread on social media, suggesting that it was caused by an act of terrorism perpetrated by Islamic extremist groups or by the so-called "yellow vests" movement.

Authors in [4] broadly identify two main approaches to fake news detection, namely content-based and context-based. *Content-based* approaches focus on shallow linguistic features such as self-reference, rhetorical structures and punctuation [5,6], and content-specific features to perform classification [7–10]. However, such features may not be consistent across media and news type [10], thus being not very useful in applications as well as for detecting fake news that are very similar to real ones. *Context-based* approaches, on the contrary, rely heavily on the social aspect of fake news, such as user reactions [11] and propagation structures [12].

In the last few years, transfer learning techniques have imposed themselves as the dominant paradigm in NLP also thanks to the introduction of the Transformer architecture and models such as BERT [13]. These systems leverage general linguistic knowledge acquired by unsupervised neural language models and are further fine-tuned to specific tasks and domains. As these models have few (and even zero) shot learning capabilities [14], the amount of data required to learn a specific task becomes much less relevant compared to their quality in describing the task itself. Therefore, the cornerstone to successfully address a downstream task now resides in developing reliable techniques to collect and label small but high-quality datasets.

Detecting fake news is a highly subjective task, and varies widely depending on when and where the fake news are produced and propagated (e.g., for specific topics or events, spread through social media). In addition, detecting fake news at the right moment is crucial. Once a piece of fake news becomes too widespread, the problem moves from its identification to its subsequent debunking. Thus, this aspect must be taken into account when creating machine learning models that deal with fake news detection, and specifically when identifying or creating reliable datasets for their training.

Two main observations can be made in this regard. On the one hand, while content-based models trained on general-purpose datasets may be suitable for easy-to-detect fake news and for performing their early detection, they may perform worse when used to discover domain-dependent and topic-dependent fake news. According to [15], adapting language models to tasks and domains is expected to improve the system performances. We can argue that also for fake news, it is preferable to train models on more focused and possibly smaller datasets referring to particular topics or events. Such focused datasets are in fact more likely to contain highly informative features about the event triggering the fake news, thereby improving their identification.

On the other hand, it is important to take into account the trade-off between the speed to create a labelled dataset and the quality of its annotation. Both ends of this spectrum have been explored in the literature. For example, some of the available datasets have been annotated with a completely automatic process, either by looking at features such as URLs or by exploiting the wisdom of the crowds via crowdsourcing experiments. While potentially noisy, this is clearly the fastest way to perform the labelling. Conversely, other datasets aimed to provide a highly reliable annotation that has been assessed and validated by expert fact-checkers, such as the one proposed in [16]. In this case, the annotation quality is superior, but at the cost of time.

To sum up, the new generation of NLP applications based on fine-tuning pre-trained language models, and the nature itself of fake news suggest that further improvements in their automatic detection may come from having small but high-quality event-dependent annotated datasets. In this paper, we propose a methodology to collect and label datasets for specific topics and events that attempts to address the trade-off between the annotation quality and the speed of the overall dataset creation process. Our approach consists of two steps: i) social media are harvested to gather data about a specific event and obtain posts and news articles that potentially propagate fake news about it; ii) the collected data are then labelled with a crowdsourced *contextualized method*, that is, annotators are provided with a small set of already known fake news to better identify and contextualize other (new) real and fake news.

In order to evaluate the quality of this annotation methodology, we use a classification model that learns the real/fake news distinction from the collected labels. Specifically, we fine-tune a Transformer language model and compare its performances with respect to i) a model based on the same architecture but trained on standard (and less polished) crowdsourced data and ii) expert-level fact-checked annotations.

We show that our method of contextualized crowdsourcing can approximate the performances of an expert human annotator. This is particularly interesting because in real-world scenarios the availability of labelled data is scarce. Thus, being able to exploit a small set of crowdsourced data to effectively learn the distinction between real and fake news on a specific topic is particularly important.

The rest of this paper is organized as follows. In Section 2, we present an overview of current fake news and rumours datasets available for research purposes. Section 3 describes our proposed methodology for collecting and effectively labelling datasets. Section 4 presents a dataset created and labelled according to the proposed method and Section 5 shows several experiments aimed at demonstrating the advantage of adopting a contextualized crowdsourcing method to create datasets suitable for training state-of-the-art classifiers. Specifically, in Section 5.1 we train and evaluate a state-of-the-art classifier on the collected dataset; Section 5.2 explores how (and to which extent) even with small amounts of data anno-

tated with the proposed methodology the classification performance improves; experiments in Section 5.3 challenge the proposed annotation methodology on a different, simulated, dataset of fake news, to further assess its effectiveness. Finally, Sections 6 and 7 discuss the results and draw some conclusions from the reported experiments.

2. Related Works

Fake news and rumour detection is a growing research area. Most works that focus on the detection problem often collect their own data for in-house validation. Therefore, this field still lacks a high-quality benchmark dataset [4]. Several challenges have to be addressed when collecting data for fake news and rumour detection, such as the identification of relevant data, their relative sparsity on social media and news websites, and current regulations concerning data access from social media [1]. Nonetheless, efforts in this direction have been made, both for fake news articles and rumours on social media.

Social media have received a lot of attention not only for automated fact-checking but in the whole area of fake news and rumour detection. For what concerns Facebook, as private user data are not available on the platform, researches focused on public profiles, such as hyper-partisan and pseudo-scientific publishers. Several datasets have been collected and labelled, with information concerning the actual news as well as *likes* and *comments* that such posts produced [17–19]. On the other hand, a lot of attention has focused on Twitter, as collecting larger-scale datasets is easier. Most of the works regarding Twitter focus specifically on the collection and annotation of data for various rumour-related tasks (e.g., rumour detection, stance classification) and on the credibility of users in the platform [16,20,21]. However, not many efforts have been specifically devoted to fake news (i.e., fake content based on news articles) on Twitter. One of the most popular repositories of fake news has been proposed in [4]. The dataset incorporates both the content of the fake news and the social context where they have spread.

In recent years, social media have also sparked researches on computational fact-checking [22]. In this case, the main focus is the collection of statements, especially from politicians and public figures, who are usually associated with a truthfulness rating as well as information regarding the context, and a brief description of the reason behind the assigned rating or related news articles [23,24]. Given the difficulties in obtaining or creating large-scale datasets with these characteristics, researchers in [25] instead proposed to alter Wikipedia sentences in order to automatically create false claims, opposed to the original ones, and provide evidence in favour or against them. More recently, a large-scale dataset of social media posts (tweets) and claims has been released in the context of the evaluation campaign CLEF 2020 with the CheckThat! task [26]. This campaign has fostered new effective methods, mostly based on modelling the semantics of the texts, both by exploiting other available datasets [27] and the sole task data [28].

The data collection strategy is also an important factor. The study in [29] describes two possible strategies for data collection. In the *top-down* approach, a-priori knowledge about rumours and fake news is used to collect relevant data about them in social media [7,30,31]. Conversely, in the *bottom-up* strategy, social media posts on breaking news are collected over a specific period of time and then analyzed by expert annotators to find and label rumours or fake news [21].

From a detection point of view, as we have already mentioned in Section 1, the approaches are divided into two main categories, namely context-based and content-based ones. Content-based approaches rely on the textual and linguistic cues of deception. For example, the use of self-reference, swear words, and negative words have been proven effective for a wide range of tasks where the intention of the author must be established. Context-based approaches are more varied and generally rely on "environmental" information, such as users' characteristics, social network propagation patterns, and reactions of other users to the news or posts. Most approaches proposed until a few years ago focused on the application of traditional machine learning algorithms, such as Support Vector Machines (SVMs) [8,9,32–34], Decision Trees or Random Forests [7,35,36], and Conditional Random Fields (CRFs) [16,37]. Nowadays, most state-of-the-art models rely on deep neural networks architectures. The main advantage of using these kinds of architectures is that they allow learning hidden representations from simpler inputs both in context and content variations [38], thus avoiding feature extraction and engineering which are prominent and time-consuming in standard machine learning approaches. Both Recurrent and Convolutional Neural Networks have been proposed to solve the task of Fake News detection, as well as ensemble and hybrid approaches [10,24,38–42]. Finally, the Transformer architecture and its transfer learning capabilities have proved to be effective for solving the fake news detection task obtaining state-of-the-art performances [43,44].

3. A new method for fake news data collection and annotation

To collect a high-quality dataset containing both real and fake news, we propose a hybrid strategy. In particular, our methodology is *bottom-up* concerning the collected data, because social media posts and related news articles are collected over a specific period of time and keywords. On the other hand, the labelling strategy is guided by a small set of a-priori fake news related to the time-span and event at hand. This set of "fake seed news" is manually identified by performing an accurate search in websites specialized in misinformation and disinformation, and provides the context for further crowdsourcing annotation, enhancing the labelling quality. In this regard, note that we do not assume the reliability of specific data sources such as international news agencies. On the contrary, our assumption in collecting the "fake seed news" was that we could consider as certainly fake only the fake news debunked by specialized services such as Snopes.[1]

[1] https://www.snopes.com

The hybrid choice is motivated by the fact that a purely bottom-up strategy with no prior knowledge requires the expertise of annotators, and is both expensive and time-consuming. To work around the problem, the proposed labelling scheme incorporates a *top-down* strategy that, starting from a small set of fake news, allows facing the annotation process with *crowdsourcing* without sacrificing data quality, which remains comparable to that obtained through annotation by experienced human users.

Fig. 1 shows the main steps of our method as a flowchart. In particular, it presents a first activity devoted to the identification of the subject of interest (e.g., a real-world event or a public figure), and two sub-processes aimed at (i) collecting data and (ii) labelling a dataset containing social media data and news about the subject of interest in a given time-span. The activities covered by the two sub-processes are detailed in the following sections.

3.1. Data Collection

Given a target event or subject, the dataset is collected starting from social media posts. Specifically, we chose to collect social media posts and news articles linked in such posts. As for the data source, we opted for Twitter since it is one of the most widely studied social networks, especially concerning rumours and fake news, and the platform offers fewer limitations than other popular social networks.

A set of posts concerning a particular subject is identified and collected by searching for specific hashtags and keywords related to the target event. From the whole dataset, we remove posts directly shared by international newspapers and news agencies (i.e., shared on their account). However, in the case that the news is retweeted or shared by other users, it is kept in the dataset and is subject to further annotation and verification. The retained tweets and news articles, which are supposed to potentially contain fake news, represent the raw data for the topic-specific dataset.

Fig. 2 summarizes the activities performed in the Data collection sub-process.

3.2. Data annotation

The key assumption behind the labelling of our dataset is that integrating contextual knowledge in a crowdsourcing experiment is essential to approximate expert human annotation. We call this approach *in-context annotation*.

We performed three annotations and rating tasks on the same data, intending to prove the added value of context. Tasks are distinguished by the information provided (or accessible) to the raters during the labelling phase.

More specifically, two of them were conducted via crowdsourcing. In the former case, annotators were provided with information on already known fake news about the subject at hand, while in the latter they were not. The possibility to access already known fake news about the event is referred to as *context* in our experiments. We employed the Prolific[2] platform and we required the annotators to be English native speakers. Each labelling experiment has been conducted independently of the other, thus annotators may or may not have participated in both tasks. The annotators were given a subset of the whole dataset and a time limit to complete the annotation and be paid for it.

The third annotation has been performed by the authors with the intent of fact-checking each piece of content. In this case, the annotator had no constraints on time and resources (i.e., the web).

The labelling methods are defined as follows:

Out-of-Context (OOC) Annotation: Given an article/post t related to the topic s, participants to a standard crowdsourcing experiment are asked to decide whether t is real or fake, without any additional supporting information. This annotation provides us with a baseline that assumes no prior knowledge of the participants regarding the event (or the fake news) being presented to them. In this experiment, annotators are expected to use their own world knowledge, either in terms of being familiar with the specific topic or with the problem of fake news (i.e., how they are written and/or how they are spread). This annotation may help us in identifying fake news that are more easily distinguishable due to their surface properties, such as typical linguistic cues or very low credibility of the information.

In-Context (IC) Annotation: Participants to a crowdsourcing experiment are asked to label a post or article as fake or real news. In this task, annotators are also given a manually selected list of known false news. The list contains the fake seed news for the subject of analysis, in the form of *source tweets/articles*, or a brief description of the fake news. This set of fake news is assumed to have been officially debunked to ascertain their falsity. Notably, these fake news may or may not be in the data presented to them, or may have a slightly different phrasing. Given an article/post t related to a topic (event or public figure) s, and a list F_s of known fake news about s, participants have to decide if t is a piece of fake news or not. By providing annotators with contextual information on specific fake posts, we expect them to better recognize the fake news in general, including more complex and harder to identify ones.

Manually fact-checked (MFC) annotation: One of the authors manually fact-checked each piece of information to assess its truthfulness by considering both the known fake news and additional sources on the web, sometimes also tracing back the news to its origin.

[2] https://www.prolific.co

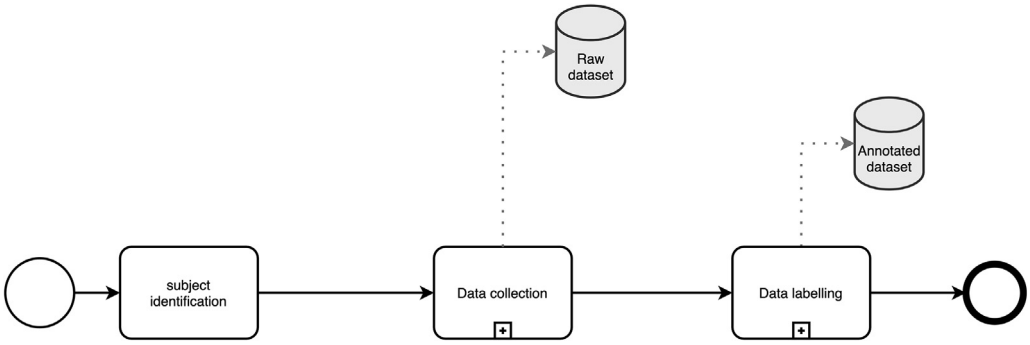

Fig. 1. Flowchart of the proposed method.

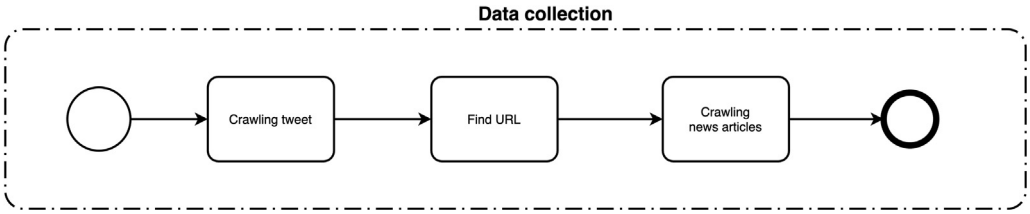

Fig. 2. Flowchart of the Data collection sub-process.

The comparison among the three approaches to collect data about fake news is also useful to shed new light on the various levels of complexity of the phenomenon. In particular, some fake news are expected to be easily identifiable from their surface characteristics while others require deeper knowledge and analysis. Different degrees of deceitfulness are therefore expected to emerge by contrasting the three methods above. For example, the fake news identified in the OOC annotation task are arguably recognizable by their surface properties or highly implausible content, while other news can be identified only by fact-checking several information or by investigating on the flow the news followed before being spread in social media.

Fig. 3 describes the whole Data labelling sub-process. Notably, the activities highlighted in blue show our proposal for a labelling scheme that leverages contextual information to provide annotators with more knowledge about the fake news (IC annotation). The remaining blocks describe the activities performed to obtain comparative data, namely the OOC annotation and the MFC one. These activities serve respectively as a baseline (i.e., a standard crowdsourcing approach) and as an optimal (i.e., a gold standard) resource to be compared with our proposed labelling scheme.

In the following section, we exemplify this methodology with the collection and labelling of fake news about the Notre-Dame Fire in 2019.

4. The Notre-Dame Fire Dataset

The Notre-Dame Fire (NDF) Dataset[3] includes news articles and tweets produced during the Notre-Dame fire of April 2019. We selected this event as our case study for two reasons: (i) it had a worldwide resonance and (ii) it was subject to the creation and spread of several fake news, thus being an appropriate candidate for evaluating our methodology. NDF is a subset of the whole dataset about the topic obtained with the data collection methodology proposed in Section 3.1. This subset was labelled via crowdsourcing on Prolific according to the methodology proposed in Section 3.2. Specifically, all contents provided in the NDF were annotated twice via crowdsourcing (out-of-context and in-context) and then manually fact-checked by one of the authors. The resulting output is a dataset provided with three levels of annotation, one for each rating task, where the final label for each item and task consists in the majority vote (for the crowdsourcing experiments) or the fact-checked decision (for the manually-fact-checked annotation).

Table 1 shows a sample of the information provided in the NDF dataset, while Tables 2 and 3 summarize the non-aggregated rating details of each task (OOC and IC) for the texts provided in Table 1. Tables 2 and 3 are connected to Table 1 via text ID. Note that the non-aggregated ratings are the absolute number of raters who actually voted for the corresponding class during the annotation process. The non-aggregated ratings are provided as part of the dataset for further analyses on fake news and their specific linguistic and contextual characteristics.

[3] The dataset is available at https://github.com/Unipisa/NDFDataset.

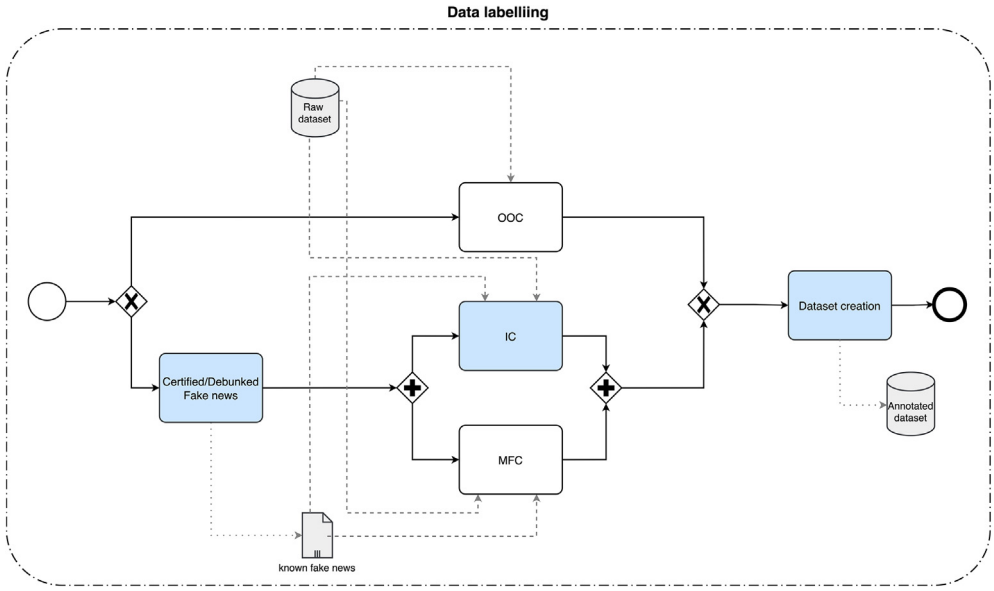

Fig. 3. Flowchart of the Data labelling sub-process.

Table 1
Sample from the NDF dataset. The provided annotation layers correspond to manually-fact-checked (MFC), in-context (IC) and out-of-context (OOC).

id	text	MFC	IC	OOC	data-type
012	Fox News host Shepard Smith abruptly cut off a French government official describing the burning of the Notre Dame cathedral after he suggested that the blaze may not have been accidental...	Fake	Fake	Fake	article
013	While the cause of the Notre Dame fire is unknown, theres a lot of noise in the catholic press about churches being targeted in France...	Fake	Fake	Real	tweet

Table 2
Non-aggregated OOC task ratings provided with the NDF dataset.

id	fake rates	real rates
012	14	6
013	7	13

Table 3
Non-aggregated IC task ratings provided with the NDF dataset.

id	fake rates	real rates
012	18	2
013	14	6

Overall, the dataset includes 568 texts, and in particular 87 articles and 481 tweets. This imbalance is due to the fact that we wanted to maintain, for the labelled subset, the same proportion between tweets and articles found in the whole collected data. However, while the data types are quite unbalanced in terms of the number of documents, they are more balanced with respect to the number of tokens, with 17,387 tweet tokens and 21,456 article tokens, for a total of 38,843.

Both crowdsourcing OOC and IC annotation tasks involved 20 participants. In OOC annotation, participants were asked to decide whether the text was propagating or not a piece of fake news without any supporting information. In the IC one, they were provided with the text (tweet or article) along with a list of already known fake news about the target event. As anticipated in Section 3, this set of fake seed news was manually identified on specialized websites. Specifically, we identified a set of 7 fake news that were spread and officially debunked early during the Notre Dame fire. In this case, participants were asked to decide whether the text was propagating one (or more) of the fake news. The two annotation methods produced rather different results in assigning the final class (fake, real) to each text piece. To assign the final class, we used the majority

vote. Given the even number of participants for each text, in some cases (10 rates per class) such an assignment is actually impossible. We can argue that these items convey a high level of uncertainty: it is not possible, with the data at hand, to decide if a particular text is spreading or not a fake news.

Finally, an additional level of annotation was added, in which one of the authors manually fact-checked each text to assess its truthfulness. In this case, both known fake news and additional sources on the web were used for fact-checking. Such level of annotation is obviously the most reliable one because the news are traced back to their origin, when possible, and therefore represent a sort of upper-level reference point to evaluate the quality of the crowdsourced data. We compared the three annotation levels on the NDF dataset and we noticed that, as expected, the class assignment varies, especially on the most difficult - hard to detect - fake news. Fig. 4 reports the class distribution across the annotation tasks. The rings, from innermost to outermost, correspond to the out-of-context (OOC), in-context (IC) and manually-fact-checked (MFC) annotation levels.

Fig. 5 shows the correlation among the three rating methods. This analysis aims to compare the correlation between the two labelling schemes (i.e., OOC and IC) and an expert human annotation (MFC). It is worth observing that the IC annotation is more strongly correlated to the MFC annotation than the OOC one. This proves the positive contribution of the seed fake news provided to the annotators during the labelling in obtaining results that are more similar to those of expert human annotation.

As noted above, the manually fact-checked annotation is arguably the most reliable one. This is because the final class was assigned by thoroughly evaluating each news with respect to its content and additional sources as well as by tracing back it to its source to assess its veracity. This allows us to evaluate how raters behave differently based on the information they had available, that is with or without some context fake news. Figs. 6 and 7 show the number of misclassified items in the two annotation experiments, with respect to the manually fact-checked data. In the OOC annotation, participants tend to confuse fake news with real ones. This behaviour is expected because some of the fake news appear credible or deceitful from a surface-based aspect. On the contrary, in the IC annotation, participants appear biased towards fake news, as they tend to confuse more real news with fake ones.

5. Evaluating the IC annotation method

In this section, we show several experiments that were carried out to test the effectiveness of our method. In Section 5.1, we present and discuss the effect of the different annotation methods on the performance of state-of-the-art classifiers using the Notre-Dame Fire dataset. We also show in Section 5.2 how even with a small amount of training data, the IC method outperforms the OOC method. Interestingly, the difference between the performances achieved by the classifiers trained with the IC and OOC methods, respectively, becomes larger as the number of training data annotated with the proposed IC method increases. In Section 5.3, we show the results of the experimentation of the IC method on another dataset, namely the PHEME dataset, which allows us to test its robustness and effectiveness on other contexts and events different from the ones used in the first dataset.

5.1. Performance of state-of-the-art classifiers

The first group of experiments has been performed on the NDF Dataset. It is aimed at evaluating how well the IC annotation method can approximate a manually fact-checked approach. In order to investigate the effect of the various annotation strategy on the performance of state-of-the-art classifiers, we fine-tune two BERT [13] classifiers, one for each type of annotated data (i.e., OOC and IC), and evaluate their performances to predict the labels provided by the manual fact-checked annotation. This allows us to evaluate how well a system trained on each specific scheme can actually predict the "correct" fact-checked labels.

For each experiment, we removed from the datasets the instances for which it was not possible to determine a membership class (i.e., the data that received even ratings). This is the reason why the two classes have a slightly different size and distribution in the two experiments. In particular, the OOC dataset contains 546 instances (177 Fake and 369 Real news) and the IC dataset contains 554 instances (216 Fake and 338 Real news). We performed stratified 10-fold cross-validation in such a way that the proportion between instances belonging to Fake and Real classes is approximately equal in all the folds.

We used a `bert-base-uncased` pre-trained model and fine-tuned it on our data. We evaluated the validation loss during training and verified that after 2 epochs the model started to overfit. We used a batch size of 8 due to computational limitations. The learning rate was set to 2e-5 following [13]. The other parameters were left to their default configuration in the huggingface[4] implementation.

Tables 4 and 5 show the performances obtained by two BERT classifiers when trained on data produced with the OOC and IC annotation paradigms. Despite the similar overall accuracy, significant differences can be observed between the two experiments for some measures, such as their ability to recognize fake news (recall on the Fake class).

Figs. 8 and 10 refer, respectively, to the OOC and the IC experiment, and contain the ROC curve as well as the AUC for each test fold.

[4] `https://huggingface.co/`

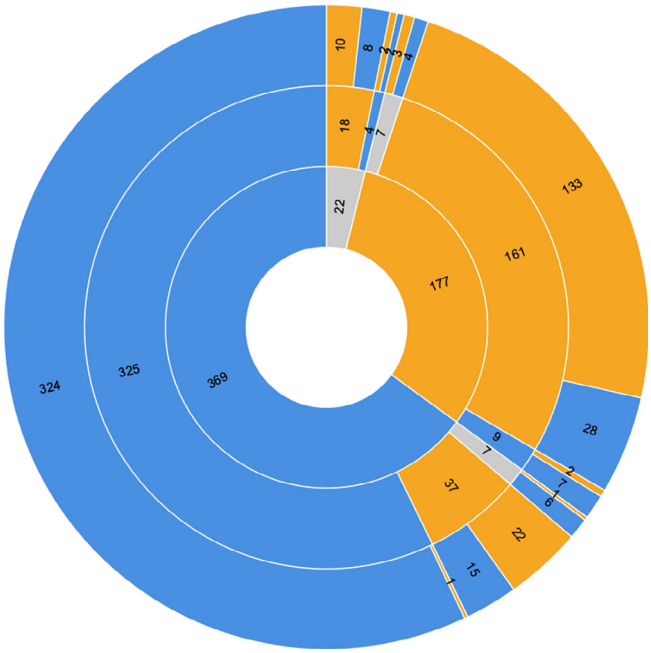

Fig. 4. Class distribution across the annotation tasks. The rings, from innermost to outermost, are the OOC, IC and MFC annotation levels. Blue regions represent real news, while orange ones represent fake news. Gray areas are data-points for which it was not possible to obtain a majority vote in a given rating task.

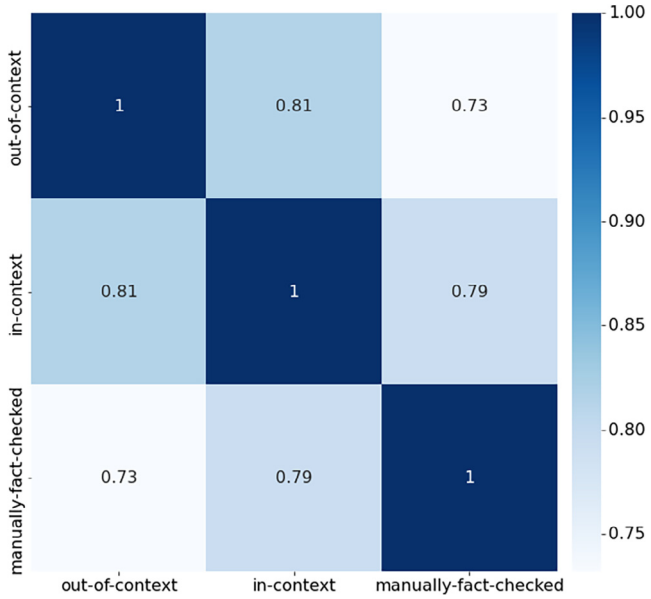

Fig. 5. Correlation among the rating tasks.

Moreover, we performed an additional experiment to evaluate the behaviour of the classifiers when they are fed with already known fake news along with the available dataset. Specifically, we added for each training fold 7 instances belonging to the Fake class and corresponding to the context provided to the raters in the IC rating task. As expected, the presence of such instances improved the performance of the classifier trained with the IC dataset given the boost attributed to the minority class. For the same reasons, we expected an improvement - albeit minor - of the classifier trained with the OOC dataset, but we noticed that, in this case, the performances degrade, probably due to the different linguistic cues and complexity of the data-points already present in the training set. Consequently, the classifier lost in generalization capability on

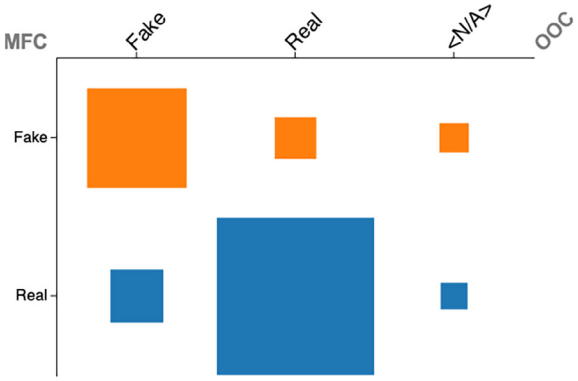

Fig. 6. OOC ratings wrt the MFC ones.

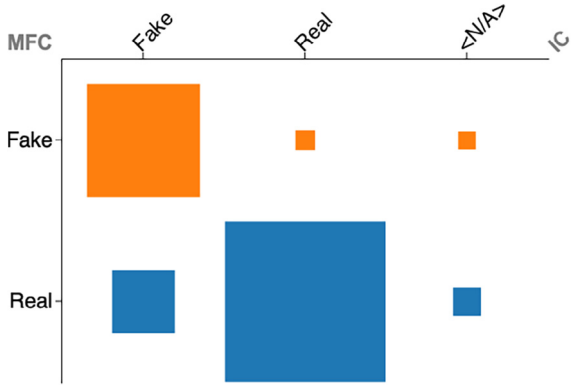

Fig. 7. IC ratings wrt the MFC ones.

Table 4
Performance obtained by the BERT classifier trained with the OOC annotated dataset.

metric	score
accuracy	**0.823**
weighted avg precision	0.831
weighted avg recall	0.823
weighted avg f1-score	0.823
macro avg precision	0.786
macro avg recall	0.783
macro avg f1-score	**0.779**
class Real precision	0.881
class Real recall	0.873
class Real f1-score	0.875
class Fake precision	0.692
class Fake recall	0.693
class Fake f1-score	0.682

the fake news class. Tables 6 and 7 show the results obtained by adding contextual fake news and the difference Δ (positive or negative) with respect to the corresponding classifier trained without adding them.

Finally, assuming that the classification agreement (i.e., the attribution of the same label across the rating tasks) can be considered as an indicator of the degree of complexity of fake news, we performed the following additional experiment. We divided our dataset into two groups, namely the highly reliable and the highly uncertain one. The news for which we had maximum confidence about the classification (i.e., the data-points with the same label for the three rating tasks) fall in the first group, as they can be considered highly reliable. The second group consists of news differently labelled across the tasks, which, contrasting with the first group, are supposed to be more uncertain and hard to classify. In the last experiment, we trained the classifier on reliable news only and we tested it on the uncertain ones. This experiment is intended as a

Table 5
Performance obtained by the BERT classifier trained with the IC annotated dataset.

metric	score
accuracy	**0.824**
weighted avg precision	0.843
weighted avg recall	0.824
weighted avg f1-score	0.828
macro avg precision	0.792
macro avg recall	0.817
macro avg f1-score	**0.797**
class Real precision	0.911
class Real recall	0.835
class Real f1-score	0.870
class Fake precision	0.673
class Fake recall	0.799
class Fake f1-score	0.725

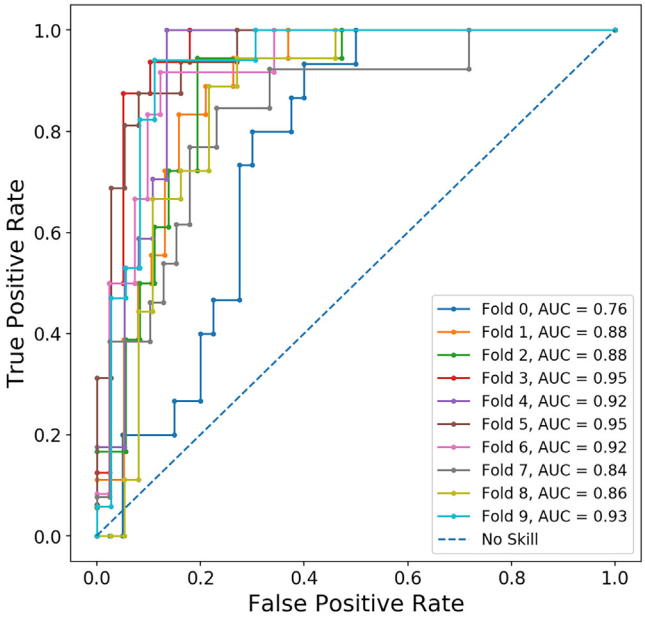

Fig. 8. ROC curves and AUCs obtained by the BERT classifier trained with the OOC annotated dataset.

sort of "stress test" in which we employ the best training set for learning the model, which is then tested on a very challenging set, composed of items for which also humans fail the classification.

We can argue that specific kinds of fake news remain difficult to classify both by human raters and automatic state-of-the-art classifiers. In fact, we can observe that in this scenario the performances deteriorate significantly, as shown in Fig. 9.

5.2. The effect of increasing training data annotated with the IC method

In this experiment, we aim to present how even a small quantity of data annotated with the IC method is enough to improve fake news identification with respect to OOC annotated data and how this improvement increases with the increasing of the size of the training set.

To this aim, we simulated a scenario in which the number of the data-points used in the training set increases over time. We split the NDF dataset into 10 nearly equal-sized folds $\{A, B, C, D, E, F, G, H, I, J\}$ and we performed a 10-fold cross-validation. For each test fold and for each $r \in [1..9]$, we trained the model with all the possible combinations of r training folds. For example, let us assume that partition A is the test fold. When we considered $r = 1$, that is, only one fold in the training set, we trained 9 models with, respectively, folds B, C, D, E, F, G, H, I, and J, and we computed the average of the performance metrics of the learned models on A. When we considered two folds in the training set (i.e. $r = 2$) we trained 36 models with, respectively, all the combinations of 2 folds in $\{B, C, D, E, F, G, H, I, J\}$ and we computed the average of the performance metrics on A. Formally, for each test fold and for each r, we trained $\frac{9!}{r!(9-r)!}$ models and we computed the average of the per-

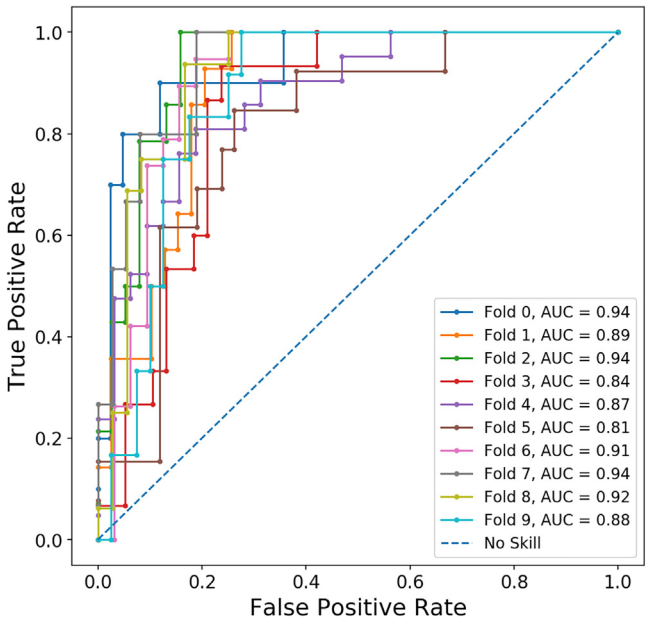

Fig. 9. BERT classifier trained on highly reliable data-points and tested on highly uncertain ones.

accuracy	0.405
weighted avg precision	0.473
weighted avg recall	0.405
weighted avg f1-score	0.397
macro avg precision	0.444
macro avg recall	0.448
macro avg f1-score	0.404
class Real precision	0.556
class Real recall	0.286
class Real f1-score	0.377
class Fake precision	0.333
class Fake recall	0.610
class Fake f1-score	0.431

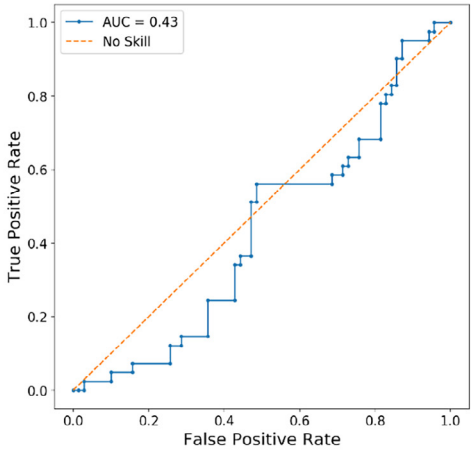

Fig. 10. ROC curves and AUCs obtained by the BERT classifier trained with the IC annotated dataset.

formance metrics for the learned models on the test fold. Finally, for each r, we averaged the performance metrics obtained in the ten test folds. Fig. 11 displays average Precision, Recall and F1-Score against increasing values of r for the OOC and IC models.

From an implementation point of view, we used overall the same strategy used in Section 5.1. We started from a `bert-base-uncased` pre-trained model and fine-tuned it on our task. Given the high number of models to be trained for the experiment, and in order to speed up the computation, we chose to exploit a more powerful computing machine that allowed us increasing the batch size to 16. All the other parameters, including the learning rate (2e-5), are the same.

Table 6
Performance obtained by the BERT classifier trained with the OOC dataset + seven known fake news.

metric	score	Δ
accuracy	**0.804**	**−1.9%**
weighted avg precision	0.818	−1.3%
weighted avg recall	0.804	−1.9%
weighted avg f1-score	0.800	−2.3%
macro avg precision	0.768	−1.8%
macro avg recall	0.766	−1.7%
macro avg f1-score	**0.752**	**−2.7%**
class Real precision	0.871	−1.0%
class Real recall	0.862	−1.1%
class Real f1-score	0.862	−1.3%
class Fake precision	0.664	−2.8%
class Fake recall	0.671	−2.2%
class Fake f1-score	0.643	−3.9%

Table 7
Performance obtained by the BERT classifier trained with the IC dataset + seven known fake news.

metric	score	Δ
accuracy	**0.849**	**+2.5%**
weighted avg precision	0.865	+2.2%
weighted avg recall	0.849	+2.5%
weighted avg f1-score	0.852	+2.4%
macro avg precision	0.817	+2.5%
macro avg recall	0.848	+3.1%
macro avg f1-score	**0.826**	**+2.9%**
class Real precision	0.933	+2.2%
class Real recall	0.847	+1.2%
class Real f1-score	0.887	+1.7%
class Fake precision	0.700	+2.7%
class Fake recall	0.849	+5.0%
class Fake f1-score	0.765	+4.0%

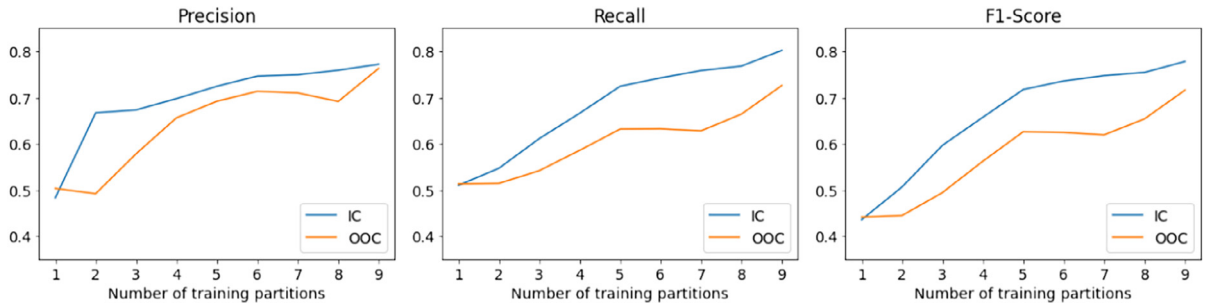

Fig. 11. Precision, Recall and F1-score against increasing values of r for the OOC and IC models.

Fig. 11 clearly shows that the IC annotation achieves better performances with a lower number of annotated examples than the OOC one. Moreover, we point out that the differences between the IC and the OOC methods tend to increase, especially in Recall and F-Score, with the increase of the number of folds considered in the training set. Overall, this experiment provides further evidence that the IC requires less training data to achieve better performances than OOC data.

5.3. Performance on a simulated dataset

In the previous sections, we have discussed the application of the IC method on the Notre-Dame Fire dataset. In this section, we investigate its behaviour on a different dataset, to test the effectiveness of the IC annotation method in other contexts and for other events. To the best of our knowledge, no datasets of fake news focused on specific real-world events are available. To simulate a dataset of fake news with this characteristic, we started from a dataset developed for a similar task, namely the popular PHEME dataset for rumour detection and veracity classification [37,45]. The PHEME dataset is suitable

for our purposes as it contains groups of rumours concerning the same real-world event. Specifically, PHEME is a dataset annotated at different levels for rumour-related tasks. It contains *Twitter conversation threads* associated with nine real-world events (Charlie Hebdo shooting, Sidney Siege hostage situation, Ferguson unrest, etc.). The *conversation* consists of a source tweet propagating a rumour and its responses that comment, support or deny the source tweet. Each tweet in the dataset is labeled either as a rumour or a non-rumour. Rumours (and associated conversation) are further grouped into *stories*, (also referred as *categories*) by expert journalists. Finally, each story is labeled as true, false, or unverified, based on if it was proven to be false, confirmed as true, or remained unverified [37,45].

To simulate our target real-world scenario (i.e., new fake news concerning real-world events), we proceeded as follows. From PHEME, we selected the tweets belonging to one of the events included in it, namely the Charlie Hebdo attack of January 7, 2015. We used the categories annotated in the dataset to infer our context (already known) fake news. We slightly modified the text of the PHEME category to construct our own context fake news. For example, given the category "Banksy drew a tribute to Charlie Hebdo featuring 3 red pencils" we provided annotators with (i) the context: "A fake news emerged in which a wall painting featuring 3 red pencils tributing Charlie Hebdo was wrongly attributed to the author Banksy." and (ii) the example Fake News: "Great picture by Banksy. So simple, yet so powerful. #JeSuisCharlie".

Overall, we supplied annotators with 9 context fake news provided with an example (fake) tweet and we asked them to annotate 262 texts (i.e., tweets). Real examples were selected from non-rumour tweets while Fake ones were selected from the list of the rumours verified as Fake. For the sake of simplicity, we decided to use a balanced dataset (i.e., 131 real news and 131 fake news) for our experiments. This choice also allowed us to measure the difference between the OOC and the IC schemes, independently of other factors such as class imbalance. We collected the judgements using Prolific and we trained a BERT [13] classifier on the annotated data to assess the method. For clarity, we refer to this annotation as PHEME-IC.

Like for the NDF dataset, we performed an additional crowdsourcing experiment providing the annotators with the same set of 262 tweets to annotate real and fake news without any context information, which is assumed as our baseline. We refer to this annotation as PHEME-OOC. In both crowdsourcing campaigns (i.e., IC and OOC), 20 annotators were asked to rate each data-point based on its veracity.

Figs. 12 and 13 show the number of misclassified items in the two annotation experiments, with respect to the PHEME gold labels. We refer to these labelling as PHEME-GOLD.

Similar to the experiments provided in Section 5.1, we used PHEME-IC and PHEME-OOC annotated data to train the classifier and we challenged the system to predict the original PHEME labels (i.e., PHEME-GOLD).

All the experiments reported in this section have been performed with the same parameters. We used a `bert-base-uncased` pre-trained model and fine-tuned it on the task data. We trained the model for 2 epochs to avoid overfitting. We used a batch size of 8 and a learning rate of 2e-5 as suggested in [13]. The other parameters were left to their default configuration in the huggingface implementation.[5]

Tables 8 and 9 and Figs. 14 and 15 report the obtained results and provide a comparison between the OOC and the IC annotation methods on the same data-points.

It is important to note that PHEME gold labels were obtained slightly differently from our manually-fact-checked annotation. In particular, to the best of our knowledge, annotators were not allowed to track the texts to their own origin to decide about their veracity. We speculate that in this case annotators started from annotated *categories* and used them to make their own decision. Therefore, in order to obtain manually-fact-checked labels similar to the ones contained in the NDF dataset, we performed an additional manual annotation in which the exploitation of additional sources on the Web was allowed. We refer to this additional annotation as PHEME-MFC.

Figs. 16 and 17 show the number of misclassified items in the two annotation experiments, namely PHEME-IC and PHEME-OOC, with respect to the PHEME-MFC ones.

Finally, we tested our PHEME-OOC and PHEME-IC classifiers on the MFC labels. The results are reported in Tables 10 and 11 and Figs. 18 and 19.

The obtained results show the inadequacy of the OOC annotation for the task. This is consistent with the results of the correlation among the various annotation schemes with respect to the PHEME-GOLD and the PHEME-MFC labels, shown in Fig. 20. Notably, the OOC and IC methods correlate very differently with both the original PHEME gold labels and an expert human annotation.

6. Discussion

Our experiments brought to light several aspects of the problem of creating and annotating datasets for fake news identification.

First of all, our experiments reveal that the task of detecting fake news is still open for both automatic systems and humans. Human abilities to recognize fake news is a crucial aspect since automatic systems are generally trained on human-rated data, thus their reliability is propagated on automatic decision systems. The experiments discussed in this paper showed that providing humans with different information during annotation impacts the quality of the labelled dataset.

[5] `https://huggingface.co/`

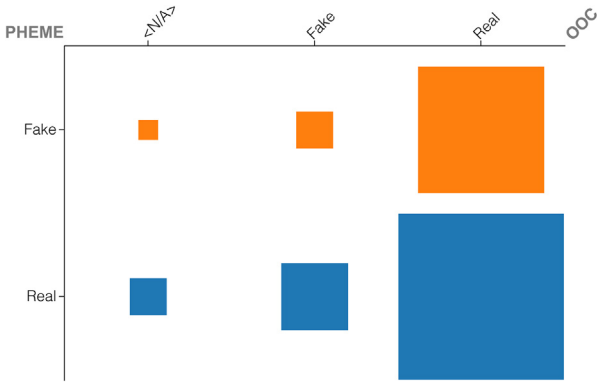

Fig. 12. PHEME-OOC ratings wrt the PHEME-GOLD.

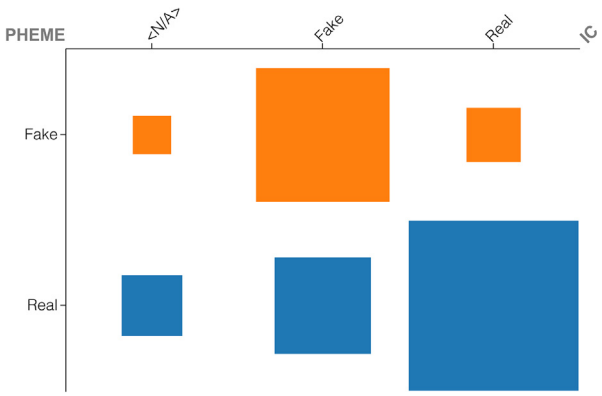

Fig. 13. PHEME-IC ratings wrt the PHEME-GOLD.

Table 8
Performance obtained by the BERT classifier trained on PHEME-OOC and tested on the PHEME-GOLD.

metric	score
accuracy	**0.642**
weighted avg precision	0.419
weighted avg recall	0.642
weighted avg f1-score	0.506
macro avg precision	0.321
macro avg recall	0.500
macro avg f1-score	**0.389**
class Real precision	0.642
class Real recall	1.000
class Real f1-score	0.779
class Fake precision	0.000
class Fake recall	0.000
class Fake f1-score	0.000

We experimented with two different ways to label the dataset, consisting of a simple standard crowdsourcing experiment where raters were asked to label a piece of content without any supporting information (OOC setting), and of a second crowdsourcing experiment in which the raters were also provided with a small set of already known fake news about the event at hand (IC setting). Both the classifiers were tested on manually fact-checked fake and real news to evaluate their capability to match expert human annotations.

It is clear that deciding if a particular piece of content is a fake or a real news is a very difficult and subjective task. On the one hand, such a decision often presupposes deep knowledge of the facts that occurred in a particular context. On the other hand, it requires the recognition of deceitfulness indicators showing the partial alteration of the content of real news.

Table 9
Performance obtained by the BERT classifier trained on PHEME-IC and tested on the PHEME-GOLD.

metric	score
accuracy	**0.727**
weighted avg precision	0.752
weighted avg recall	0.727
weighted avg f1-score	0.724
macro avg precision	0.717
macro avg recall	0.705
macro avg f1-score	**0.695**
class Real precision	0.813
class Real recall	0.753
class Real f1-score	0.768
class Fake precision	0.620
class Fake recall	0.658
class Fake f1-score	0.623

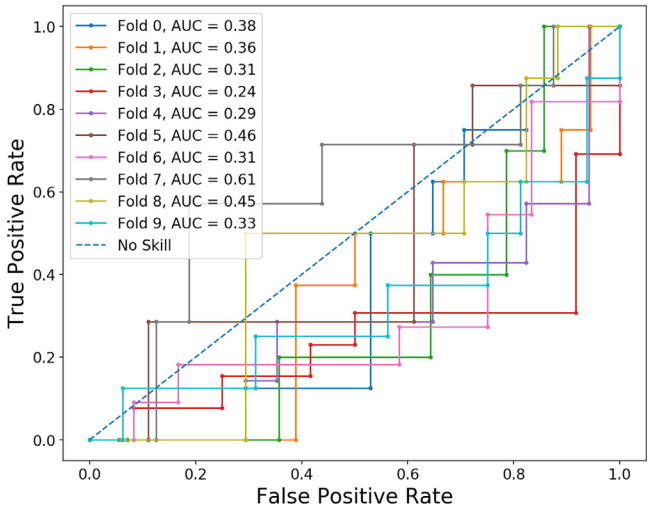

Fig. 14. ROC curves and AUCs obtained by the BERT classifier trained on PHEME-OOC and tested on the PHEME-GOLD.

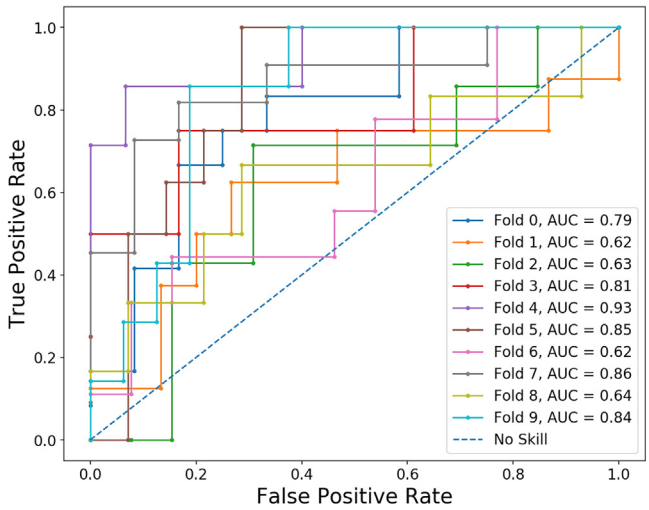

Fig. 15. ROC curves and AUCs obtained by the BERT classifier trained on PHEME-IC and tested on the PHEME-GOLD.

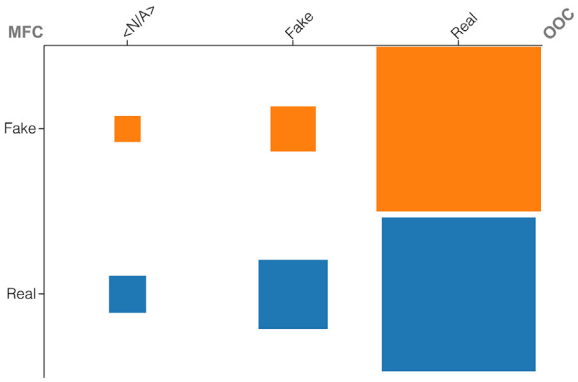

Fig. 16. PHEME-OOC ratings wrt the PHEME-MFC.

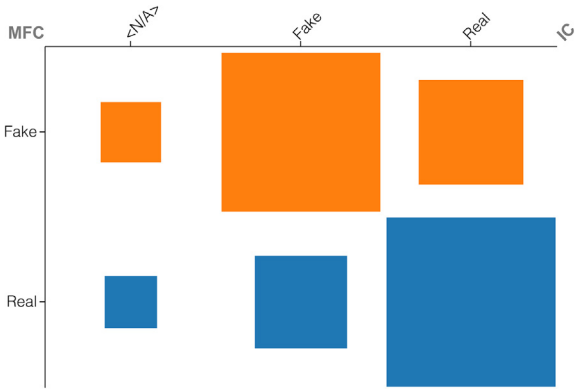

Fig. 17. PHEME-IC ratings wrt the PHEME-MFC.

Table 10
Performance obtained by the BERT classifier trained on PHEME-OOC and tested on PHEME-MFC.

metric	score
accuracy	**0.481**
weighted avg precision	0.240
weighted avg recall	0.481
weighted avg f1-score	0.318
macro avg precision	0.240
macro avg recall	0.500
macro avg f1-score	**0.322**
class Real precision	0.481
class Real recall	1.000
class Real f1-score	0.645
class Fake precision	0.000
class Fake recall	0.000
class Fake f1-score	0.000

We are well aware that the ideal case would be to train automatic classifiers on gold data labelled by experienced fact-checkers, but we also know that real-world applications can hardly afford such type of in-depth manual annotation of entire datasets in a short time. However, we can argue that providing a small amount of contextual information to the annotators at rating time allows us to obtain a satisfactory approximation of the ideal setting.

The first interesting insight obtained by analysing our results is the fact that the labelling methodology is a key factor when trying to match the quality of a manually annotated dataset containing an expert's evaluation. We were able to show that by providing contextual information to the annotators, the resulting labels were more helpful for the classifier in actually predicting the correct label for unseen, and manually fact-checked data. This may be due to two factors.

Table 11
Performance obtained by the BERT classifier trained on PHEME-IC and tested on PHEME-MFC.

metric	score
accuracy	**0.743**
weighted avg precision	0.781
weighted avg recall	0.743
weighted avg f1-score	0.734
macro avg precision	0.774
macro avg recall	0.739
macro avg f1-score	**0.727**
class Real precision	0.710
class Real recall	0.844
class Real f1-score	0.757
class Fake precision	0.838
class Fake recall	0.635
class Fake f1-score	0.698

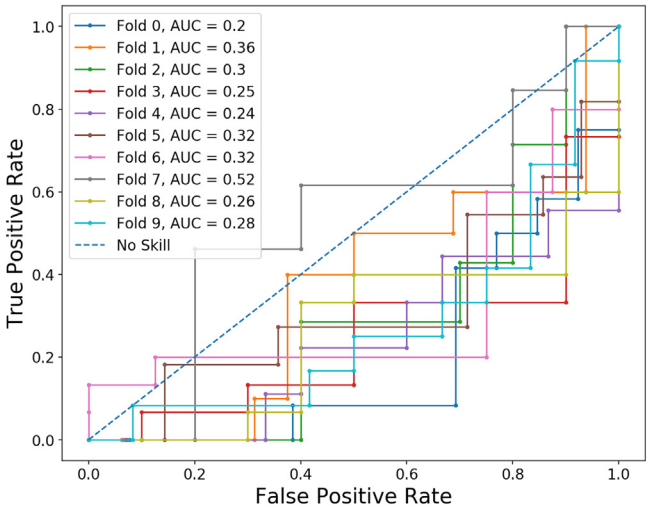

Fig. 18. ROC curves and AUCs obtained by the BERT classifier trained on PHEME-OOC and tested on the manually-fact-checked labels.

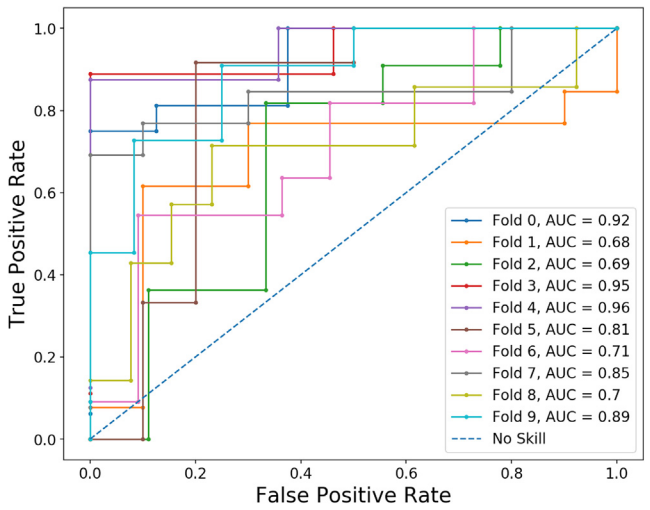

Fig. 19. ROC curves and AUCs obtained by the BERT classifier trained on PHEME-OOC and tested on the manually-fact-checked labels.

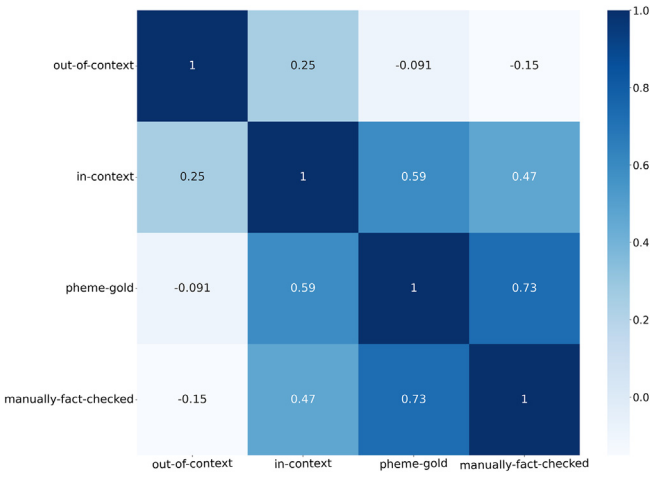

Fig. 20. Correlation among the rating tasks on the adapted version of PHEME.

First, the annotation scheme where raters are given more contextual information on potential fake news helped them in producing more consistent labels with respect to the dataset labelled by the expert. This is confirmed by looking at the correlation matrices between the various annotation methods. In the case of the NDF dataset, we can see a difference in terms of correlation between the IC and the OOC method of 6% with respect to manually-fact-checked data (see Section 4, Fig. 5). This difference becomes much more impressive on the PHEME dataset. In this case the difference in correlation between the IC and OOC methods with respect to PHEME-GOLD labels is 68%, while it is 62% with respect to PHEME-MFC ones (see Section 5.3, Fig. 20).

Second, as annotators are aware of the fact that the data they are analysing may contain fake news similar to the ones provided to them, they may be biased to label them as such. This, in turn, has the effect of boosting the performances of the classifier on the "Fake" class. This aspect is evident in both the NDF and the PHEME datasets. Interestingly, such bias also exists when comparing this classifier with a similar one trained on manually fact-checked data on the NDF dataset. In this case, we observe that the "biased" classifier performs better than the one trained on "correctly" labelled data. We can hypothesize that a classifier learned on a vast amount of "correctly" labelled data would yield even better results. Nevertheless, the realistic situation in which a rather small topic-focused set of data can be collected and rapidly labelled to train a classifier can benefit from the context-induced bias to improve the capability of recognizing the fake (minority) class. In addition to this, as we aim also to empower users to detect news that may need further verification (i.e., potentially fake ones), biasing the classifier towards improving the recall on this specific class may eventually be an efficient and effective choice in real-world scenarios.

Moreover, since we did not observe this situation in the PHEME dataset, this effect may also be due to data imbalance. In PHEME, in fact, where fake and real news are perfectly balanced, training the classifier on manually-fact-checked data is better than training it on IC labelled data. We speculate that the effect of the bias is much more prominent in unbalanced datasets like real-world ones are.

Concerning the experiments shown in Section 5.1, we can make more in-depth observations. Comparing the results obtained from the two classifiers trained on crowdsourcing data (see Tables 4 and 5 and Figs. 8 and 10), we noticed that the IC model has a much higher ability to distinguish fake from real news, with an overall higher macro-average F-Score (+2%), and a marked better ability to recognize fake news, with a higher Recall (+10%) on the Fake class.

The second experiment of this subsection was oriented to simulate the real-world scenario in which the classifiers are fed with few, already known data, in our case regarding 7 already known fake news (see Tables 6 and 7). As expected, adding such data-points improved the performance of the IC classifier given the boost attributed to the minority class. However, no improvement has been observed by adding them to the OOC classifier, which instead deteriorated in predicting manually-fact-checked data. This fact is probably due to the diverse linguistic cues and complexity of the data-points already present in the OOC training set, namely the ones annotated without knowing anything about already established fake news. We can argue that the class attributed by raters to these data-points in the OOC crowdsourcing setting is mostly based on their surface characteristics and general credibility and not on other more content-based aspects of the phenomenon, such as their similarity with already known, intentionally distorted, and false news.

Finally, we challenged the classifiers on predicting the veracity of complex pieces of text. In this case we observed that some fake news remain almost impossible to classify for both humans and automatic systems. We isolated a set of news for which we did not find agreement among the labelling schemes, starting from the assumption that such news are the most complex of the dataset. This set of news has been used as a test set while the remaining instances have been used for training a new classifier. This experiment pointed out that, even if we train a classifier on "perfect" data (with complete agreement

among all our labelling schemes, including the manually-fact-checked one), it is very difficult to recognize particularly complex fake news because of their linguistic and semantic features. Interestingly, those complex news may also be considered as a list of hard-to-detect fake news that emerge as a byproduct of our annotation experiments. In other words, it could be argued that the fake news labelled as fake in all the labelling tasks are the easiest to detect based on their textual characteristics, while those labelled differently across the tasks have a higher level of complexity and deceitfulness.

Beyond particularly complex cases, the overall results obtained for fake news detection are in line with the state-of-the-art, despite they were obtained by fine-tuning a BERT classifier on a relatively small amount of data. Several strategies contributed to this fact. First, the proposed approach is natively domain-adapted because it starts from the collection of the news related to a particular target event or topic. Second, the support provided by the context is twofold. On the one hand, it helps raters to produce labels that are more correlated with the ones labelled by an expert. On the other hand, it is helpful when added as a set of additional instances in the training phase.

Moreover, the experiments aimed at comparing the number of data required to achieve satisfactory results with both OOC and IC methods revealed that the use of an in-context annotation allows classifiers to obtain better performances for all volumes of data considered. Actually, the gap between the performances increases as the data size increases (see Fig. 11). This shows us that the method is more effective also when the volume of available data is limited.

Finally, by applying the proposed methodology to a different dataset we were able to assess its robustness and generality. We are aware that the PHEME dataset is targeted to a different (albeit similar) task, and that its adaptation to our needs may have introduced some noise. Nonetheless, evaluating the differences between the PHEME-OOC and PHEME-IC annotations definitely showed us the effectiveness of the IC method, taking into account the clear difficulties encountered by raters in discerning real from fake news. The PHEME-OOC annotation strongly diverges from the original gold labels (see Fig. 16), making the resulting classifier unfeasible for the task (see Tables 8 and 10). Conversely, by providing raters with a small number of already verified fake news about the event for the PHEME-IC annotation, we were able to achieve acceptable results (see Tables 9 and 11), in line with the experiments conducted on the NDF dataset. This is also proven by considering the performances achieved by the classifier trained on the NDF dataset with comparable amounts of data, as shown in Fig. 11.

7. Conclusions

We have introduced a new methodology for collecting and labelling datasets of fake and real news. The proposed method is oriented towards real applications potentially requiring a rapid domain adaptation. In these cases, the goal is to quickly label high-quality data on particular events or subjects to be used to adapt existing models or to create new ones. In this paper we focused on the second aspect, assessing the prediction ability of models built on a rather small domain-dependent dataset.

The collection strategy starts from social media and gathers both social media data and newspaper articles linked to them. The approach has been evaluated on a dataset consisting of real and fake news concerning the Notre Dame Fire of 2019.

Given such data, a new annotation method is proposed, which uses few contextual data consisting of already known fake news about the topic with the aim of both accelerating the labelling and improving the quality of the labelled data. Moreover, contextual data have also proven to be effective as training instances. To evaluate how much the labelling scheme impacts the final results, a state-of-the-art classifier has been trained and evaluated for its ability to match gold, manually-fact-checked data. On the one hand, our study shows the limitation of using standard crowdsourcing settings to collect data about highly subjective tasks. On the other hand, it proves that such limitations can be overcome by providing all the raters with adequate contextual knowledge about the subject.

We have shown that already with a small amount of training data annotated with the IC method, the accuracy of the classifiers increases with respect to the training data annotated with standard OOC crowdsourcing. Furthermore, we have pointed out how this difference in performance becomes larger as training data annotated with the proposed method increase.

We have also experimented the IC method on another dataset, namely the PHEME dataset. We aimed to verify its robustness and effectiveness on other contexts and events. The obtained results have proved that the proposed method allows considerable improvement with respect to the standard OOC crowdsourcing in different scenarios.

Moreover, our work has confirmed that further research is needed to deal with particularly complex fake news, which remain obscure and difficult to detect for both humans and automatic detection systems. We hope that future research on these facets will be enabled by releasing the NDF dataset. Albeit being small and focused on a single subject, the presence of several annotation layers will enable different types of experiments and serve for exploring linguistic and propagation characteristics of the data. Furthermore, comparing the label distribution of the data among the annotation layers will shed light on a crucial aspect of fake news, namely their different degrees of deceitfulness and duplicity.

The methodology for improved data annotation we propose is independent of the specific domain and task. In particular, it may be applied to tasks where rating can be impacted by the actual knowledge of facts and/or by a strong subjective component. In this paper, we addressed the former aspect, and we showed that annotators provided with factual knowledge are better able to label the data for fake news detection. We believe that the latter aspect is also very interesting and complex. In

future works, we will focus on jointly modelling subjective tasks and factual-oriented ones. In this context, it would be crucial to study the interaction between factual knowledge and subjectivity, for example for the identification of hate speech that is generated by fake news. We expect that our methodology can be used widely in the scientific community, thus enabling the creation of a public repository with different datasets annotated as we have proposed in the paper. The repository would allow developing and experimenting with novel approaches to fake news detection and related problems, thus contributing to solve or at least mitigate the negative impact and social effects of misinformation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. This work was partially supported by the University of Pisa in the context of the project "Event Extraction for Fake News Detection" in the framework of the MIT-Unipi program, and by the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence).

References

[1] A. Bondielli, F. Marcelloni, A survey on fake news and rumour detection techniques, Information Sciences 497 (2019) 38–55.
[2] H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election, Tech. rep., National Bureau of Economic Research (2017).
[3] A. Zubiaga, M. Liakata, R. Procter, K. Bontcheva, P. Tolmie, Towards detecting rumours in social media, in: AAAI Workshop: AI for Cities, 2015, pp. 35–41.
[4] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD Explorations Newsletter 19 (1) (2017) 22–36.
[5] E.J. Briscoe, D.S. Appling, H. Hayes, Cues to deception in social media communications, in: Proceedings of the 2014 47th Hawaii International Conference on System Sciences (HICSS), IEEE, 2014, pp. 1435–1443.
[6] V.L. Rubin, T. Lukoianova, Truth and deception at the rhetorical structure level, Journal of the Association for, Information Science and Technology 66 (5) (2015) 905–917.
[7] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the 20th international conference on World Wide Web, ACM, Hyderabad, India, 2011, pp. 675–684.
[8] V. Rubin, N. Conroy, Y. Chen, S. Cornwell, Fake news or truth? using satirical cues to detect potentially misleading news, in: Proceedings of the NAACL-CADD2016 Second Workshop on Computational Approaches to Deception Detection, San Diego, California, USA, 2016, pp. 7–17.
[9] B.D. Horne, S. Adali, This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News, arXiv e-prints (2017) arXiv:1703.09398.
[10] N. Ruchansky, S. Seo, Y. Liu, Csi: A hybrid deep model for fake news detection, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 797–806.
[11] V. Qazvinian, E. Rosengren, D.R. Radev, Q. Mei, Rumor has it: Identifying misinformation in microblogs, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Edinburgh, Scotland, UK, 2011, pp. 1589–1599.
[12] J. Ma, W. Gao, K.-F. Wong, Detect rumors in microblog posts using propagation structure via kernel learning, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol. 1, 2017, pp. 708–717.
[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.
[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates Inc, 2020, pp. 1877–1901.
[15] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N.A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 8342–8360.
[16] A. Zubiaga, M. Liakata, R. Procter, Learning Reporting Dynamics during Breaking News for Rumour Detection in Social Media, arXiv e-prints (2016) arXiv:1610.07363.
[17] M. Potthast, S. Köpsel, B. Stein, M. Hagen, Clickbait detection, in: European Conference on Information Retrieval, Springer, 2016, pp. 810–817.
[18] E. Tacchini, G. Ballarin, M.L. Della Vedova, S. Moret, L. de Alfaro, Some like it hoax: Automated fake news detection in social networks, CoRR abs/1704.07506.
[19] G.C. Santia, J.R. Williams, Buzzface: A news veracity dataset with facebook user commentary and egos, in: Proceedings of the 12th International AAAI Conference on Web and Social Media, 2018, pp. 531–540.
[20] T. Mitra, E. Gilbert, Credbank: A large-scale social media corpus with associated credibility annotations, in: Proceedings of the 9th International AAAI Conference on Web and Social Media, 2015, pp. 258–267.
[21] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, A. Zubiaga, Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, 2017, pp. 69–76.
[22] Y. Wu, P.K. Agarwal, C. Li, J. Yang, C. Yu, Toward computational fact-checking, Proceedings of the VLDB Endowment (PVLDB) 7 (7) (2014) 589–600.
[23] A. Vlachos, S. Riedel, Fact checking: Task definition and dataset construction, in: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Association for Computational Linguistics, 2014, pp. 18–22.
[24] W.Y. Wang, "liar, liar pants on fire": A new benchmark dataset for fake news detection, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, 2017, pp. 422–426.
[25] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 809–819.

[26] S. Shaar, A. Nikolov, N. Babulkov, F. Alam, A. Barrón-Cedeño, T. Elsayed, M. Hasanain, R. Suwaileh, F. Haouari, G. Da San Martino, P. Nakov, Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media, 2020.
[27] M. Bouziane, H. Perrin, A. Cluzeau, J. Mardas, A. Sadeq, Team buster. ai at checkthat! 2020: Insights and recommendations to improve fact-checking, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, 2020.
[28] L.C. Passaro, A. Bondielli, A. Lenci, F. Marcelloni, Unipi-nle at checkthat! 2020: Approaching fact checking from a sentence similarity perspective through the lens of transformers, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, 2020.
[29] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, R. Procter, Detection and resolution of rumours in social media: A survey, ACM Comput. Surv. 51 (2) (2018) 32:1–32:36.
[30] J. Ma, W. Gao, Z. Wei, Y. Lu, K.-F. Wong, Detect rumors using time series of social context information on microblogging websites, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM, Melbourne, VIC, Australia, 2015, pp. 1751–1754.
[31] S. Vosoughi, M. Mohsenvand, D. Roy, Rumor gauge: predicting the veracity of rumors on twitter, ACM Transactions on Knowledge Discovery from Data (TKDD) 11 (4) (2017) 50.
[32] Y. Qin, D. Wurzer, V. Lavrenko, C. Tang, Spotting rumors via novelty detection, CoRR abs/1611.06322.
[33] F. Yang, Y. Liu, X. Yu, M. Yang, Automatic detection of rumor on sina weibo, in: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, ACM, 2012, p. 13.
[34] K. Wu, S. Yang, K.Q. Zhu, False rumors detection on sina weibo by propagation structures, in: Proceedings of the 2015 IEEE 31st International Conference on Data Engineering (ICDE), IEEE, 2015, pp. 651–662.
[35] G. Giasemidis, C. Singleton, I. Agrafiotis, J.R. Nurse, A. Pilgrim, C. Willis, D.V. Greetham, Determining the veracity of rumours on twitter, in, in: International Conference on Social Informatics, Springer, 2016, pp. 185–205.
[36] Z. Zhao, P. Resnick, Q. Mei, Enquiring minds: Early detection of rumors in social media from enquiry posts, in, in: Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Florence, Italy, 2015, pp. 1395–1405.
[37] A. Zubiaga, M. Liakata, R. Procter, Exploiting context for rumour detection in social media, in: G.L. Ciampaglia, A. Mashhadi, T. Yasseri (Eds.), Social Informatics: 9th International Conference, Springer International Publishing, Cham, 2017, pp. 109–123.
[38] J. Ma, W. Gao, P. Mitra, S. Kwon, B.J. Jansen, K.-F. Wong, M. Cha, Detecting rumors from microblogs with recurrent neural networks, in: IJCAI'16 Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 2016, pp. 3818–3824.
[39] Y.-C. Chen, Z.-Y. Liu, H.-Y. Kao, Ikm at semeval-2017 task 8: Convolutional neural networks for stance detection and rumor verification, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017, pp. 465–469.
[40] F. Yu, Q. Liu, S. Wu, L. Wang, T. Tan, A convolutional approach for misinformation identification, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, AAAI Press, 2017, pp. 3901–3907.
[41] S. Volkova, K. Shaffer, J.Y. Jang, N. Hodas, Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vol. 2, 2017, pp. 647–653.
[42] A. Zubiaga, E. Kochkina, M. Liakata, R. Procter, M. Lukasik, K. Bontcheva, T. Cohn, I. Augenstein, Discourse-aware rumour stance classification in social media using sequential classifiers, Information Processing & Management 54 (2018) 273–290.
[43] V. Slovikovskaya, G. Attardi, Transfer learning from transformers to fake news challenge stance detection (FNC-1) task, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 1211–1218.
[44] M. Qazi, M.U.S. Khan, M. Ali, Detection of fake news using transformer model, in: 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2020, pp. 1–6.
[45] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, P. Tolmie, Analysing how people orient to and spread rumours in social media by looking at conversational threads, PLOS ONE 11 (3) (2016) 1–29.