

# WorthIt: Check-worthiness Estimation of Italian Social Media Posts

Agnese Daffara<sup>1,2</sup>, Alan Ramponi<sup>3</sup> and Sara Tonelli<sup>3,\*</sup>

<sup>1</sup>Institute for Natural Language Processing, University of Stuttgart – Stuttgart, Germany

<sup>2</sup>Department of Humanities, University of Pavia – Pavia, Italy

<sup>3</sup>Digital Humanities group, Fondazione Bruno Kessler – Trento, Italy

## Abstract

Check-worthiness estimation is the first and a paramount task in the automated fact-checking pipeline. It allows professional fact-checkers to cope with the increasing amount of mis/disinformative textual content being published online by prioritizing claims that are factual/verifiable and worthy of verification. Despite the long tradition of check-worthiness estimation in NLP, there is currently a lack of annotated resources and associated methods for Italian. Moreover, current datasets typically cover a single topic and focus on a limited time frame, affecting models' generalizability on out-of-distribution data. To fill these gaps, in this paper we introduce WORTHIT, the first annotated dataset for factuality/verifiability and check-worthiness estimation of Italian social media posts that covers public discourse on migration, climate change, and public health issues across a large time period of six years. We describe the dataset creation in detail and conduct thorough experimentation with the WORTHIT dataset using a wide array of encoder- and decoder-based models. Our results show that fine-tuning monolingual encoder-based models in a multi-task setting provides the best overall performance and that decoder-based models in a few-shot setup still struggle in capturing the relation between factuality/verifiability and check-worthiness. We release our dataset, code, and associated materials to the research community.

## Keywords

Automated fact-checking, check-worthiness estimation, factual/verifiable claim detection, resources and evaluation

## 1. Introduction

Given the unprecedented amount of mis/disinformation spreading online, assisting fact-checkers in their everyday work by automatizing some of their tasks is becoming of paramount importance. The identification of content that is worthy of verification – i.e., *check-worthiness estimation*, also referred to as *check-worthy claim detection* – represents the first stage in the fact-checking pipeline [1] insofar as it allows professional fact-checkers to reduce the screening efforts of content that is not worth of attention, therefore focusing on the verification of potentially false or misleading information.

According to Nakov et al. [2], a claim is deemed check-worthy and calls for the attention of a fact-checker if it “is likely to be false, is of public interest, and/or appears to be harmful”, also being not “easy to fact-check by a layperson” (e.g., “The capital of Italy is Rome”). A check-worthy claim is both *factual* and *verifiable* [2, 3], i.e., it presents an “assertion about the world that is checkable” [4], namely it “state[s] a definition, mention[s] a quantity in the present or in the past, make[s] a verifiable



**Figure 1:** Example of social media posts classified according to their factuality/verifiability (FV) and check-worthiness (CW) and their relation to the verification process by a fact-checker.

prediction of the future, reference[s] laws, procedures, and rules of operation, discuss[es] images or videos, [or] state[s] correlation or causation” [2]. In other words, if a claim is factual and verifiable, it is possible to determine its check-worthiness based on whether it is relevant and may potentially have a broader impact on the general public [5] (see examples in Figure 1).

Check-worthy claim detection<sup>1</sup> has become a well-established task in NLP since the introduction of the first CheckThat! evaluation campaign [6]. However, despite the progress and the coverage of multiple languages in the past CheckThat! editions, no dataset or task for check-worthiness estimation specifically for the Italian language has been considered so far. Moreover, current

The repository is publicly available on GitHub at: <https://github.com/dhfbk/worthit>.

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 – 26, 2025, Cagliari, Italy

\*Corresponding author.

agnese.daffara@ims.uni-stuttgart.de (A. Daffara);  
alramponi@fbk.eu (A. Ramponi); satonelli@fbk.eu (S. Tonelli)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

<sup>1</sup>In this paper, we refer to the task as “check-worthy claim detection” or “check-worthiness estimation” interchangeably.

datasets for check-worthiness estimation in other languages mostly focus on COVID-19 issues and consist of posts that were drawn from a relatively small time period (e.g., one year and three months [2]), affecting out-of-distribution generalization of models [7].

**Contributions** In this paper, we address the aforementioned gaps by developing **WORTHIT**, the first annotated dataset for factuality/verifiability and check-worthiness estimation for Italian, and by conducting extensive experiments with encoder- and decoder-based models. **WORTHIT** covers public discourse from Twitter on migration, climate change, and public health issues over a large time frame of six years. The full dataset was annotated by two expert annotators which discussed the cases of disagreement to resolve annotation errors (e.g., due to attention drops) while keeping genuine annotation divergences (e.g., due to different interpretations), in line with recent work advocating the importance of considering human label variation in subjective tasks [8, 9, 10, 11, 12, *inter alia*]. We fine-tune a wide array of monolingual and multilingual encoder-based models in single- and multi-task learning settings, and experiment with four decoder-based models that include Italian in pretraining data in a few-shot setup after a careful selection of representative examples. Results show that multi-task fine-tuning of encoder-based models provides the best performance, and that decoder-based models – with or without annotation guidelines in the prompt, either in Italian or English – still struggle in tackling the task effectively, even when provided with information about the factuality/verifiability of the post.

## 2. Related Work

Check-worthy claim detection is a popular task within the NLP community mostly thanks to the series of CheckThat! shared tasks organized by the CLEF initiative.<sup>2</sup> Indeed, check-worthy claim detection is the only task that has been proposed at all seven CheckThat! editions [6, 13, 14, 15, 2, 16, 17]. Several datasets for training check-worthiness estimation models in different languages have been created and released, starting from English and Arabic at CheckThat! 2018 [6] to Arabic, Bulgarian, Dutch, English, Spanish, and Turkish in later editions [2, 16, 17]. Besides CheckThat! datasets, additional resources have been developed over the years, mainly focused on specific events like COVID-19 [18] or political news [19]. English is the most represented language for check-worthiness estimation, but the scientific community has recently started to focus on the development of resources for other languages too, since check-worthy claims can refer to events that are relevant

only for areas in which a given language is spoken. In this respect, Italian represents an exception because, to our knowledge, no dataset for check-worthiness estimation in this language has been developed so far. Recently the Check-IT! dataset [20] has been created, which however contains only fact-checked (i.e. check-worthy) claims. Likewise, the FEVER-IT dataset [21] is a translation into Italian of the widely-used FEVER dataset [22], and contains only claims to be verified against textual sources. In this work, we address this gap by presenting the novel **WORTHIT** dataset, which covers a previously overlooked language for the task of check-worthiness estimation. The dataset has been carefully sampled across topics and time for better models’ generalizability, since past works have shown that the performance of automated fact-checking drops under domain shift [23]. The **WORTHIT** dataset has also been fully annotated by two raters to value human label variation [8].

Concerning the methods for check-worthiness estimation, state-of-the-art results in the CheckThat! evaluation campaigns are mostly based on fine-tuned encoder-based models such as BERT, RoBERTa, and DistilBERT [24, 25, 26] and language-specific variants [16], often combined with data augmentation [27] and ensembling strategies. Recently, large language models (LLMs) have been started to be used for the task, showing promising performance. For instance, the best performing system on English at CheckThat! 2024 [17] fine-tuned Llama-2-7B on the provided training data and then leveraged prompts generated by ChatGPT for check-worthy claim detection [28]. However, previous works do not leverage the synergies between factuality/verifiability and check-worthiness, albeit being strictly related tasks. Our work makes a step towards this goal by fine-tuning encoder-based models in a multi-task learning setting and experimenting with sequential prompting using decoder-based models.

## 3. WORTHIT Dataset

In this section, we describe the dataset creation process, from data collection (Section 3.1) to data annotation (Section 3.2). We then present data statistics (Section 3.3).

### 3.1. Data Collection

We collect social media posts pertaining to migration, climate change, and public health issues using the Twitter APIs.<sup>3</sup> To mitigate temporal bias in the dataset, we focus on a large time frame of six full years (from 2017-01-01 to 2022-12-31) and retain messages in Italian about the aforementioned topics by using a manually curated list of over 400 keywords derived from reliable glossaries and

<sup>2</sup><https://www.clef-initiative.eu/>.

<sup>3</sup>Tweets were retrieved in 02/2023 when the APIs for research purposes were still available for free.

scientific manuals (see Appendix A). Following Nakov et al. [2], we further filter out posts containing  $\leq 5$  tokens<sup>4</sup> and sort the remaining messages by their sum of likes and retweets. We then select the top- $k$  ( $k = 10$ ) posts exhibiting the highest number of likes and retweets for each month and topic subset, therefore focusing on the messages with the highest impact to the society while simultaneously mitigating topic and temporal biases. We further account for the potential presence of authors’ writing style biases that can occur when many posts authored by the same users are included in the dataset: we therefore retain only the most impactful post authored by the same user in each data subset. Overall, we collect 2,160 posts evenly distributed across topics (i.e., 720 for each topic) and time periods (i.e., 360 for each year) for factuality/verifiability and check-worthiness annotation. All posts have been then anonymized by replacing user mentions, URLs, email addresses, and phone numbers with placeholders (i.e., [USER], [URL], [EMAIL], and [PHONE], respectively) and newline characters (i.e.,  $\backslash n$  and  $\backslash r$ ) have been replaced with single spaces.

### 3.2. Data Annotation

Each post has been annotated with two labels, namely *i*) one denoting whether the content of the post is factual/verifiable – either YES or NO – and *ii*) one indicating its check-worthiness – with labels in a 5-point Likert scale: DEFINITELY YES, PROBABLY YES, NEITHER YES NOR NO, PROBABLY NO, or DEFINITELY NO. It is worth noting that, as opposed to determining factuality/verifiability, estimating check-worthiness is a partly subjective task. This motivates us to create WORTHIT with parallel labels by the annotators on all posts so that future studies on human label variation can be conducted. The annotation guidelines closely follow the ones used in Check-That! shared tasks and are provided in Appendix B.

**Annotators** Annotation was conducted by two expert annotators. Both annotators are native speakers of Italian and have naturally been exposed to public discourse on migration, climate change, and public health in the Italian context. They identify themselves as a woman and a man, with age ranges 20–30 and 30–40. They have a background in linguistics and natural language processing and conducted annotation as part of their work.

**Annotation process** Annotators were provided with annotation guidelines for determining the factuality/verifiability and check-worthiness of social media posts (Appendix B). After conducting a pilot annotation phase on a small subset of the messages, annotators discussed

the cases in which their annotations diverged, and consolidated the guidelines by specifying how to deal with special cases (e.g., in the presence of reported speech; see Appendix B). Then, they both labeled the full set of posts in four rounds of annotation. Each round involved a discussion phase aimed at resolving annotation errors (e.g., due to attention slips) while keeping instances exhibiting genuine disagreement (e.g., different interpretations). This makes WORTHIT the first check-worthiness estimation dataset that goes beyond the “single ground truth” assumption in subjective annotation.

**Inter-annotator agreement** We computed the inter-annotator agreement (IAA) on the full dataset for both factuality/verifiability and check-worthiness using Krippendorff’s alpha ( $\alpha$ ) [29]. We obtain 0.8322 for factuality/verifiability and 0.6909 for check-worthiness. As expected, albeit substantial, the IAA for check-worthiness is lower than that for factuality/verifiability due to the genuine disagreement that we retain on purpose.

### 3.3. Data Analysis and Statistics

WORTHIT comprises 2,160 posts distributed across topics and time periods as shown in Figure 2, in which we also highlight the overlap in posts with FAINA [30], a previously released dataset for fine-grained fallacy detection. Specifically, WORTHIT includes the same posts from 2019 to 2022 that are in FAINA and further includes messages from 2017 and 2018 time periods. This opens opportunities for studying the interplay between check-worthiness and fallacious argumentation in future work as well as investigations on human label variation, especially because annotators are the same for both datasets.

	In WORTHIT only		In both WORTHIT and FAINA			
Migration	120	120	120	120	120	120
Climate change	120	120	120	120	120	120
Public health	120	120	120	120	120	120
	2017	2018	2019	2020	2021	2022

**Figure 2:** Distribution of posts in WORTHIT across topics and time periods and overlap in posts with the FAINA dataset.

Overall, social media posts in WORTHIT have an average token length of 38.6 and the full dataset comprises 83,315 tokens, of which 28,562, 26,667, and 28,086 are part of migration, climate change, and public health posts, respectively. In Table 1 we summarize the annotation statistics for factuality/verifiability and check-worthiness for both annotators ( $A_1$  and  $A_2$ ). While 1,413–1,432 posts (65.4%–66.3%) are considered as factual/verifiable

<sup>4</sup>Computed using the `it_core_news_sm` spaCy model (v3.5).

**Table 1**

Factuality/verifiability (FV) and check-worthiness (CW) label statistics in WORTHIT across annotators ( $A_1$  and  $A_2$ ). Check-worthiness labels from left to right are: DEFINITELY NO, PROBABLY NO, NEITHER YES NOR NO, PROBABLY YES, and DEFINITELY YES.

		ANNOTATOR $A_1$				
		NO		YES		
FV		747 (34.6%)	1,413 (65.4%)			
CW		43 (2.0%)	342 (15.8%)	17 (0.8%)	807 (37.4%)	204 (9.4%)
		← NO			YES →	
		ANNOTATOR $A_2$				
		NO		YES		
FV		728 (33.7%)	1,432 (66.3%)			
CW		145 (6.7%)	380 (17.6%)	123 (5.7%)	574 (26.6%)	210 (9.7%)
		← NO			YES →	

by annotators, we stress that the check-worthiness of a post can be estimated only if the post itself is deemed as factual/verifiable. Indeed, only 1,011 (46.8%) and 784 (36.3%) posts over the total are classified as check-worthy (i.e., either with the label PROBABLY YES or DEFINITELY YES) by  $A_1$  and  $A_2$ , respectively. We observe that the overall statistics for factuality/verifiability are similar among annotators, while those for check-worthiness, as expected, vary more. Specifically,  $A_1$  appears to have been more inclined to assign clear-cut check-worthiness scores, whereas  $A_2$  distributed its ratings more across the scale. While in our experiments (Section 4) we do not directly leverage this information, our dataset is released to the community with disaggregated labels using the full 5-point Likert scale to encourage work on fine-grained check-worthiness estimation and human label variation.

## 4. Experiments

We conduct experiments on check-worthiness estimation with both encoder- and decoder-based models using the newly-introduced WORTHIT dataset. In this section, we thoroughly detail our experimental setup (Section 4.1) and the model variant and prompt selection process (Section 4.2). Then, we present test set results (Section 4.3).

### 4.1. Experimental Setup

**Task setup** We cast the check-worthiness task as a binary classification problem and consider factuality/verifiability as auxiliary information that can be leveraged by models to improve performance on the task. Given that each post has annotations provided by all annotators (i.e.,  $A_1$  and  $A_2$ ) for both factuality/verifiability and check-worthiness, for the purpose of the experiments

we consider a post as factual/verifiable if both annotators agree that the instance is such (i.e., labeling it as YES), whereas we consider a post as check-worthy if both annotators label the instance as either PROBABLY YES or DEFINITELY YES. This ensures that these posts are check-worthy with high likelihood. As a result, we obtain 1,341 (62.1%) and 819 (37.9%) factual/verifiable and non factual/verifiable posts, and 739 (34.2%) and 1,421 (65.8%) check-worthy and non check-worthy posts, respectively.

**Data splits** We divide WORTHIT into  $k$  training and test sets using  $k$ -fold cross-validation ( $k = 5$ ) preserving the label distribution across splits. For development, we rely on the training portions only and further divide them into training and development sets for the purpose of model variant and prompt selection (Section 4.2). Specifically, for encoder-based models we split them into five 80%/20% training/development sets, while for decoder-based models we divide them into two equal parts: the first half is used for selecting examples for few-shot prompting, while the second half serves as development set. All texts were lowercased for the purpose of the experiments.

**Models** For the experiments with encoder-based models, we use four monolingual models specifically trained on Italian data, namely ALBERTo [31],<sup>5</sup> UmBERTo [32],<sup>6</sup> and dbmdz’s Italian BERT models [33] in their base<sup>7</sup> and xxl<sup>8</sup> versions (henceforth referred to as BERT-it base and BERT-it xxl). Moreover, we employ widespread multilingual models that include Italian in pretraining data, namely mBERT [34]<sup>9</sup> and XLM-RoBERTa [35].<sup>10</sup> For fine-tuning, we use the MaChAmp toolkit (v0.4.2) [36] and select the best hyperparameter configuration based on average Pos F<sub>1</sub> score on the development sets (Appendix C). As regards decoder-based models, we choose two Italian and two multilingual models, all instruction-tuned. Specifically, we select LLaMantino-3-ANITA-8B [37]<sup>11</sup> and Minerva-7B [38]<sup>12</sup> as monolingual models, while we use Qwen2.5-7B [39]<sup>13</sup> and Llama3.1-8B [40]<sup>14</sup> as multilingual models. We choose these models because they are widely used, freely available, and do not require very large computational resources that could be impractical in real-world scenarios. Predicted labels are extracted from models’ outputs using regular expressions. If no

<sup>5</sup>Version: m-polignano-uniba/bert\_uncased\_L-12\_H-768\_A-12\_italian\_alb3rt0

<sup>6</sup>Version: Musixmatch/umberto-commoncrawl-cased-v1

<sup>7</sup>Version: dbmdz/bert-base-italian-uncased

<sup>8</sup>Version: dbmdz/bert-base-italian-xxl-uncased

<sup>9</sup>Version: google-bert/bert-base-multilingual-cased

<sup>10</sup>Version: FacebookAI/xlm-roberta-base

<sup>11</sup>Version: swap-uniba/LLaMantino-3-ANITA-8B-Inst-DPO-ITA

<sup>12</sup>Version: sapienzanlp/Minerva-7B-instruct-v1.0

<sup>13</sup>Version: Qwen/Qwen2.5-7B-Instruct

<sup>14</sup>Version: meta-llama/Meta-Llama-3.1-8B-Instruct



matching label is found in the output,<sup>15</sup> the response is recorded as “unknown”. Hyperparameter details are in Appendix C. Overall, we employ six encoder-based and four decoder-based models, for a total of ten models.

**Prompts and example sets** For decoder-based models, we design prompts in two languages (Italian and English) with or without annotation guidelines, leading to four different prompt configurations: Italian with guidelines (*it\_g*), Italian without guidelines (*it\_ng*), English with guidelines (*en\_g*), and English without guidelines (*en\_ng*). All models are prompted in a few-shot setup with five carefully-selected examples of posts and associated labels (Section 4.2).<sup>16</sup> All prompts are in Appendix D.

### Multi-task fine-tuning and sequential prompting

We hypothesize that factuality/verifiability information can help to predict the check-worthiness of a post. We thus design different fine-tuning and prompting settings for encoder- and decoder-based models, respectively, to test this hypothesis. Specifically, for encoder-based models we compare a standard *SINGLE TASK* approach (i.e., fine-tuning a model with check-worthiness labels only) with an approach that leverages both factuality/verifiability and check-worthiness information in a *MULTI-TASK* learning framework (i.e., using check-worthiness as a main task and factuality/verifiability as an auxiliary task with different task loss weights  $\lambda_{fv}$  and  $\lambda_{cw}$ ; see Appendix C). We compute the multi-task learning loss as  $L = \sum_t \lambda_t L_t$ , where  $L_t$  is the loss for the task  $t$ , i.e., either factuality/verifiability (*fv*) or check-worthiness (*cw*), and  $\lambda_t$  is the weight given to the task. For decoder-based models, we instead test a standard setting in which the models are prompted directly for check-worthiness (*NOT SEQ*) and a two-step sequential prompting approach (*SEQ*) (prompt are in Appendix D). In the latter case, the model is firstly instructed to classify the post based on its factuality/verifiability, then the output label is incorporated into a prompt which instructs the model to assess the check-worthiness of the same post.

**Evaluation metrics** We use the  $F_1$  score for the positive check-worthy class (Pos  $F_1$ ) as our main metric, in line with previous work on check-worthy claim detection [2, 16, 17, *inter alia*]. For completeness, we also report positive precision and recall scores (Pos Prec and Pos Rec, respectively), as well as accuracy (Acc) for test set results. Since encoder-based models provide confidence scores for the output labels, we also compute mean

average precision (mAP) scores for them to get additional insights on performance when ranking posts by check-worthiness. Moreover, for decoder-based models we include the number of “unknown” outputs (i.e., those not matching a label in the label set) to assess their ability to follow the instructions.

## 4.2. Model Variant and Prompt Selection

We select the most promising setting (i.e., model variant, set of few-shot examples, and prompt configuration) based on average Pos  $F_1$  score on the development sets. While for encoder-based approaches the model selection was mainly a matter of tuning hyperparameter values (see Section 4.1 and additional details in Appendix C), for decoder-based models this involved the selection of the most promising set of examples as well as the prompt configuration (i.e., language and guidelines).

**Few-shot example set selection** We create five different sets of few-shot examples (i.e., post texts and associated labels) by diversifying them across topics and annotation combinations for factuality/verifiability and check-worthiness, focusing on examples that are similar to those that are discussed in the annotation guidelines. Each set is drawn from one of the five training splits used during development and contains five examples. Table 2 reports the composition of each set with respect to topics and annotations. To select the most promising example set to be used in the test phase, we prompt all decoder-based models with these example sets. In Table 2 we also report the Pos  $F_1$  obtained by using each example set, averaged on all models, development sets, and prompt configurations across *SEQ* and *NOT SEQ* settings (calculated over a total of 138,400 data points).<sup>17</sup> Example set #1 leads to the highest average Pos  $F_1$  score and also exhibits the smallest standard deviation (Table 2); therefore, we select this set for the test phase (refer to Appendix D for post texts and labels included in the example set). It is worth noting that this is the only set that does not include any post annotated as factual/verifiable but not check-worthy (+-), suggesting that models may learn more effectively from examples that are either both factual/verifiable and check-worthy or neither. In Table 3, we report the percentages of factuality/verifiability and check-worthiness label combinations outputted by models when prompted using each example set over all the possible configurations in the *SEQ* setting (69,200 data points). We observe that even if the sets have different

<sup>15</sup> Allowed labels for factuality/verifiability: {factual, fattuale, not[\_]factual, non[\_]fattuale}; allowed labels for check-worthiness: {check[\_]worthy, not[\_]check-worthy, non[\_]check-worthy}.

<sup>16</sup> Testing a smaller/larger number of examples is left for future work.

<sup>17</sup> Each development split for decoder-based models consists of 865 examples (i.e., 50% of the training portion; see Section 4.1). Therefore, we have 865 outputs per development set ( $5 \times$ )  $\rightarrow$  4,325 outputs per model’s configuration ( $4 \times$ )  $\rightarrow$  17,300 outputs per model ( $4 \times$ )  $\rightarrow$  69,200 outputs per setting ( $2 \times$ )  $\rightarrow$  138,400 outputs in total.

**Table 2**

Composition of the five example sets assessed in the development phase in terms of covered topics (Mi: migration, CC: climate change, PH: public health) and label combinations (++: factual/verifiable and check-worthy; --: not factual/verifiable and not check-worthy; +-: factual/verifiable and not check-worthy). Pos F<sub>1</sub> scores are averaged on all models, development sets, and prompt configurations across SEQ and NOT SEQ settings. The score for the best example set is in **bold**.

Set	Topic			FV/cw			Pos F <sub>1</sub> score
	Mi	CC	PH	++	--	+-	
#1	2	2	1	3	2	0	<b>0.68237</b> ±0.10
#2	2	1	2	3	1	1	0.63348±0.11
#3	2	2	1	2	2	1	0.68000±0.11
#4	2	2	1	2	2	1	0.59510±0.14
#5	1	2	2	1	2	2	0.59459±0.12

**Table 3**

Percentages of factuality/verifiability (FV) and check-worthiness (cw) label combinations outputted by models when prompted using each example set over all configurations in the SEQ setting (computed on the development sets).

FV/cw	#1	#2	#3	#4	#5
++	16.64%	13.72%	15.91%	8.01%	9.04%
+-	5.18%	3.40%	4.82%	2.95%	7.35%
-+	43.76%	43.80%	46.92%	43.57%	35.71%
--	34.43%	39.08%	32.35%	45.46%	47.89%

distributions of label combinations, this does not influence significantly the distribution of the labels generated by models: in all cases, models frequently produce an invalid pair -FV +cw, while they tend to avoid the opposite one (i.e., +FV -cw).

**Best prompt selection** To select the prompts for the test phase, we compare average Pos F<sub>1</sub> scores on the development splits obtained by all decoder-based models when prompted with `it_g`, `it_ng`, `en_g`, and `en_ng` configurations (21,625 data points for each configuration)<sup>18</sup> in both SEQ and NOT SEQ settings. Results are in Table 4. All the best performing models do not use guidelines; therefore, we decide not to include guidelines in the prompts in further experiments. We keep both English and Italian prompt versions for the test phase, as some models perform better with Italian (particularly Minerva). We also observe that the best results in the SEQ setting are overall higher than in the direct check-worthiness task (i.e., NOT SEQ). We keep both settings for testing to better highlight performance differences.

<sup>18</sup>865 outputs per development set (5×) → 4,325 outputs per example set (5×) → 21,625 outputs in total.

**Table 4**

Development set results for check-worthiness estimation using decoder-based models in SEQ and NOT SEQ settings, split by prompt configuration. We report F<sub>1</sub> scores for the positive *check-worthy* class (Pos F<sub>1</sub>; *main metric*). Results are averaged across  $k = 5$  development splits and example sets and include standard deviations. The best setting for each model is underlined and the best overall result is in **bold**.

Model	Setting	Config	Pos F <sub>1</sub>
LlaMAntino-3-ANITA-8B	NOT SEQ	<code>en_ng</code>	0.6759±0.06
		<code>it_ng</code>	0.6645±0.05
		<code>en_g</code>	0.5728±0.11
	SEQ	<code>it_g</code>	0.5400±0.11
		<code>en_ng</code>	<u>0.7552</u> ±0.03
		<code>it_ng</code>	0.6658±0.06
		<code>en_g</code>	0.6802±0.07
		<code>it_g</code>	0.5523±0.10
	Minerva-7B	<code>en_ng</code>	0.5724±0.04
		<code>it_ng</code>	0.6338±0.03
Qwen2.5-7B	NOT SEQ	<code>en_g</code>	0.6331±0.02
		<code>it_g</code>	0.6451±0.02
		<code>en_ng</code>	0.4825±0.04
	SEQ	<code>it_ng</code>	<u>0.6832</u> ±0.02
		<code>en_g</code>	0.5541±0.04
		<code>it_g</code>	0.6695±0.01
		<code>en_ng</code>	0.6709±0.13
	Llama3.1-8B	<code>it_ng</code>	0.5976±0.13
		<code>en_g</code>	0.6397±0.10
		<code>it_g</code>	0.6149±0.09
		<code>en_ng</code>	<u>0.7419</u> ±0.07
Llama3.1-8B	NOT SEQ	<code>it_ng</code>	0.6157±0.11
		<code>en_g</code>	0.6456±0.11
		<code>it_g</code>	0.5901±0.10
	SEQ	<code>en_ng</code>	<b>0.7805</b> ±0.01
		<code>it_ng</code>	0.7661±0.02
		<code>en_g</code>	0.7434±0.04
		<code>it_g</code>	0.6932±0.06
	NOT SEQ	<code>en_ng</code>	0.7515±0.06
		<code>it_ng</code>	0.7682±0.01
		<code>en_g</code>	0.7374±0.05
	SEQ	<code>it_g</code>	0.7211±0.05

### 4.3. Results

We compute the results for the selected configurations of encoder- and decoder-based models across the  $k = 5$  test splits, presenting average scores and standard deviations across the applicable metrics as detailed in Section 4.1.

**Encoder-based models** Results for encoder-based models are shown in Table 5. We observe that using factuality/verifiability as an auxiliary task in a MULTI-TASK learning framework helps to improve the Pos F<sub>1</sub> performance across all models. The best scores are obtained by BERT-it xxl, followed by UmBERTo and BERT-it base, all fine-tuned in a MULTI-TASK setting. Specifically, BERT-it xxl fine-tuned using both factuality/verifiability and check-worthiness information achieves a Pos F<sub>1</sub> score

**Table 5**

Test set results for check-worthiness estimation using fine-tuned encoder-based models in SINGLE TASK and MULTI-TASK settings. We report  $F_1$  scores for the positive *check-worthy* class (Pos  $F_1$ ; *main metric*), along with positive precision (Pos Prec) and recall (Pos Rec) scores. We further report mean average precision (mAP; *secondary metric*) scores and accuracy (Acc) scores. Results are averaged across  $k = 5$  test splits and include standard deviations. For main and secondary metrics, the best setting for each model is underlined and the best overall result is in **bold**.

Model	Setting	Pos $F_1$	Pos Prec	Pos Rec	mAP	Acc
AlBERTo	SINGLE TASK	0.7039 $\pm$ 0.03	0.6397 $\pm$ 0.04	0.7848 $\pm$ 0.03	0.7563 $\pm$ 0.04	0.7731 $\pm$ 0.03
	MULTI-TASK	<u>0.7107</u> $\pm$ 0.02	0.6258 $\pm$ 0.03	0.8240 $\pm$ 0.02	<u>0.7713</u> $\pm$ 0.03	0.7699 $\pm$ 0.02
UmBERTo	SINGLE TASK	0.7247 $\pm$ 0.02	0.6413 $\pm$ 0.03	0.8349 $\pm$ 0.03	<u>0.7974</u> $\pm$ 0.04	0.7829 $\pm$ 0.02
	MULTI-TASK	<u>0.7277</u> $\pm$ 0.02	0.6432 $\pm$ 0.03	0.8403 $\pm$ 0.04	0.7958 $\pm$ 0.04	0.7847 $\pm$ 0.02
BERT-it base	SINGLE TASK	0.7121 $\pm$ 0.02	0.6694 $\pm$ 0.04	0.7646 $\pm$ 0.05	0.7770 $\pm$ 0.04	0.7884 $\pm$ 0.02
	MULTI-TASK	<u>0.7146</u> $\pm$ 0.03	0.6698 $\pm$ 0.04	0.7687 $\pm$ 0.04	<u>0.7805</u> $\pm$ 0.03	0.7898 $\pm$ 0.02
BERT-it xxl	SINGLE TASK	0.7332 $\pm$ 0.02	0.7066 $\pm$ 0.03	0.7646 $\pm$ 0.04	0.8054 $\pm$ 0.03	0.8097 $\pm$ 0.02
	MULTI-TASK	<u>0.7473</u> $\pm$ 0.02	0.7017 $\pm$ 0.02	0.8010 $\pm$ 0.04	<b>0.8095</b> $\pm$ 0.03	0.8148 $\pm$ 0.01
mBERT	SINGLE TASK	0.6767 $\pm$ 0.03	0.5831 $\pm$ 0.04	0.8105 $\pm$ 0.05	0.7384 $\pm$ 0.03	0.7347 $\pm$ 0.03
	MULTI-TASK	<u>0.6828</u> $\pm$ 0.03	0.5904 $\pm$ 0.04	0.8132 $\pm$ 0.04	<u>0.7496</u> $\pm$ 0.04	0.7407 $\pm$ 0.03
XLM-RoBERTa	SINGLE TASK	0.7014 $\pm$ 0.02	0.6293 $\pm$ 0.02	0.7929 $\pm$ 0.02	0.7441 $\pm$ 0.03	0.7690 $\pm$ 0.01
	MULTI-TASK	<u>0.7138</u> $\pm$ 0.02	0.6313 $\pm$ 0.03	0.8241 $\pm$ 0.03	<u>0.7621</u> $\pm$ 0.02	0.7736 $\pm$ 0.02

**Table 6**

Test set results for check-worthiness estimation using decoder-based models in SEQ and NOT SEQ settings, split by prompt language. We report  $F_1$  scores for the positive *check-worthy* class (Pos  $F_1$ ; *main metric*), positive precision (Pos Prec), recall (Pos Rec), and accuracy (Acc) scores, along with “unknown” outputs. Results are averaged across  $k = 5$  test splits and include standard deviations. For the main metric, the best setting for each model is underlined and the best overall result is in **bold**.

Model	Setting	Lang	Pos $F_1$	Pos Prec	Pos Rec	Acc	Unknown
LlaMAntino-3-ANITA-8B	NOT SEQ	en	0.6556 $\pm$ 0.03	0.5959 $\pm$ 0.03	0.7294 $\pm$ 0.04	0.7380 $\pm$ 0.02	0
		it	0.6409 $\pm$ 0.02	0.5661 $\pm$ 0.02	0.7402 $\pm$ 0.03	0.7162 $\pm$ 0.02	0
	SEQ	en	<b>0.6771</b> $\pm$ 0.02	0.5980 $\pm$ 0.02	0.7808 $\pm$ 0.02	0.7449 $\pm$ 0.02	0
		it	0.6111 $\pm$ 0.03	0.5654 $\pm$ 0.03	0.6671 $\pm$ 0.04	0.7093 $\pm$ 0.03	0
Minerva-7B	NOT SEQ	en	0.3506 $\pm$ 0.01	0.2706 $\pm$ 0.01	0.4980 $\pm$ 0.01	0.3690 $\pm$ 0.01	81 $\pm$ 2
		it	0.3629 $\pm$ 0.01	0.2610 $\pm$ 0.01	0.5954 $\pm$ 0.02	0.2847 $\pm$ 0.01	112 $\pm$ 8
	SEQ	en	0.2944 $\pm$ 0.00	0.2357 $\pm$ 0.00	0.3924 $\pm$ 0.01	0.3565 $\pm$ 0.01	127 $\pm$ 8
		it	<u>0.4442</u> $\pm$ 0.02	0.3075 $\pm$ 0.01	0.7997 $\pm$ 0.04	0.3157 $\pm$ 0.01	58 $\pm$ 4
Qwen2.5-7B	NOT SEQ	en	0.5917 $\pm$ 0.02	0.4286 $\pm$ 0.01	0.9553 $\pm$ 0.02	0.5486 $\pm$ 0.02	0
		it	<u>0.6273</u> $\pm$ 0.01	0.4697 $\pm$ 0.01	0.9445 $\pm$ 0.01	0.6157 $\pm$ 0.02	0
	SEQ	en	0.5885 $\pm$ 0.01	0.4211 $\pm$ 0.01	0.9770 $\pm$ 0.00	0.5324 $\pm$ 0.01	0
		it	0.6247 $\pm$ 0.02	0.5016 $\pm$ 0.02	0.8281 $\pm$ 0.03	0.6597 $\pm$ 0.02	0
Llama3.1-8B	NOT SEQ	en	0.5470 $\pm$ 0.00	0.3780 $\pm$ 0.00	0.9892 $\pm$ 0.01	0.4394 $\pm$ 0.01	0
		it	<u>0.5616</u> $\pm$ 0.01	0.3955 $\pm$ 0.01	0.9689 $\pm$ 0.02	0.4824 $\pm$ 0.02	0
	SEQ	en	0.5585 $\pm$ 0.01	0.3895 $\pm$ 0.01	0.9864 $\pm$ 0.01	0.4662 $\pm$ 0.01	0
		it	0.5584 $\pm$ 0.01	0.3929 $\pm$ 0.01	0.9648 $\pm$ 0.01	0.4778 $\pm$ 0.01	0

of 0.7473 (+1.41 points increase compared to the SINGLE TASK version) and a mAP score of 0.8095 on the check-worthiness estimation task. Notably, XLM-RoBERTa in a MULTI-TASK setting shows only -3.35 points than the best BERT-it xxl configuration in terms of Pos  $F_1$  score, despite being pretrained on a mixture of languages. It also outperforms AlBERTo in the MULTI-TASK setup and obtains comparable results in the SINGLE TASK setting. This

suggests that XLM-RoBERTa can be a viable approach for multilingual check-worthiness estimation.

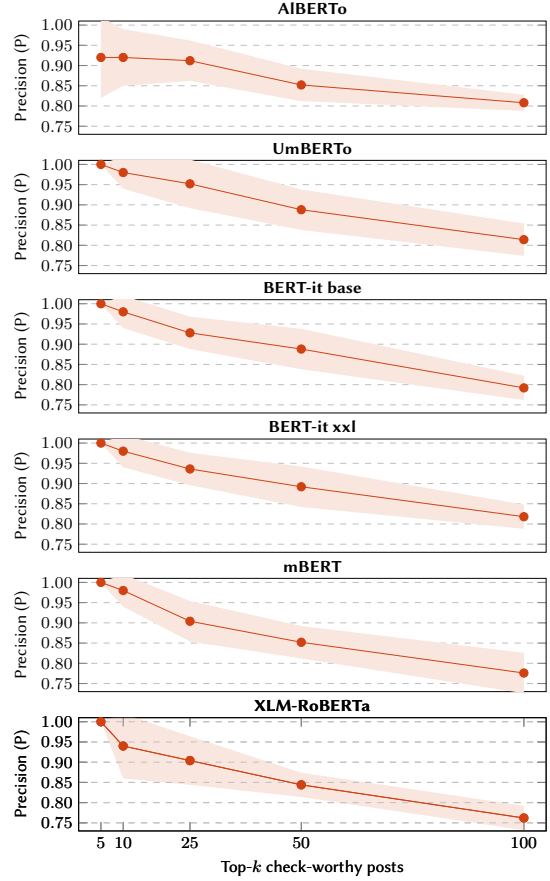
**Decoder-based models** Results are presented in Table 6. Decoder-based models in a few-shot setup perform slightly worse on average than fine-tuned encoder-based models, but still achieve competitive results. Moreover, three models perform better when prompted in Italian.

Notably, LLaMAntino-3-ANITA-8B – despite being pre-trained on Italian data – performs better with English prompts and achieves the highest score in the SEQ setting (i.e., 0.6771 Pos F<sub>1</sub> score). The two Italian models, LLaMAntino-3-ANITA-8B and Minerva-7B, reach the best results in the SEQ setup, while the multilingual models Qwen2.5-7B and Llama3.1-8B perform better when directly prompted for check-worthiness (i.e., in the NOT SEQ setup). Overall, factuality and verifiability information do not seem to significantly aid decoder-based models in predicting check-worthiness, as they are unable to leverage this information effectively (see Section 5 for an in-depth analysis). The lowest performance is observed with Minerva-7B, which is also the only model to produce “unknown” outputs – up to an average of 127 “unknown” labels when prompted in English in the SEQ setting.

## 5. Analysis and Discussion

**Ranking of posts by check-worthiness** Aggregate check-worthiness estimation scores (e.g., Pos F<sub>1</sub>) give a useful picture of models’ performance; however, knowing how the models *rank* the posts by check-worthiness is paramount for fact-checkers since they can only screen a limited number of posts in their daily work (say,  $k$ ). In Figure 3, we report the ratio of posts correctly classified as check-worthy within the top- $k$  recommended check-worthy posts ( $P@k$ ) by all encoder-based models,<sup>19</sup> with  $k \in \{5, 10, 25, 50, 100\}$ . We observe that  $P@k$  is in the range of 0.90–0.95 and 0.80–0.85 points on average when the posts’ screening budget is set to  $k = 25$  and  $k = 100$ , respectively. This indicates that these models can help fact-checkers in their daily routine.

**Relationship between fv and cw** To assess whether decoder-based models capture the relationship between factuality/verifiability and check-worthiness, we analyzed their outputs in the SEQ setup. Figure 4 shows the frequencies of the four possible combinations of labels both in the models’ outputs (i.e., +FV +CW, +FV -CW, -FV +CW, and -FV -CW; calculated over 8,650 data points) and in the manual annotations (2,160 data points). The most frequent label combination in the models’ outputs is +FV +CW, accounting for more than half of the predictions for Minerva-7B and Llama3.1-8B, reaching 66.2% for the latter. Interestingly, the second most frequent combination is -FV +CW: we consider this as problematic, because non-factual or non-verifiable posts should not be classified as check-worthy. This suggests that decoder-based models do not grasp this correlation and instead

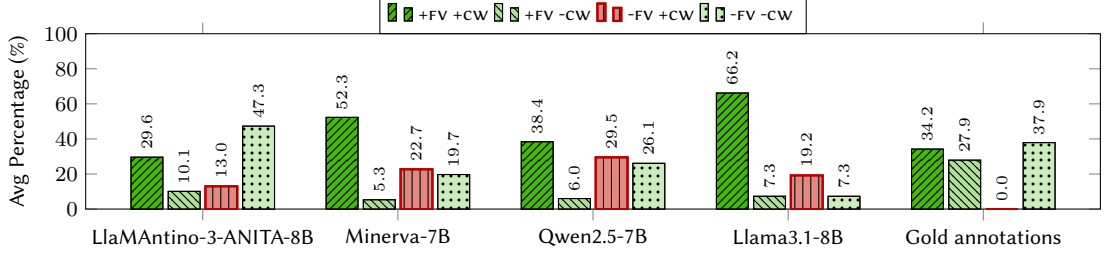


**Figure 3:**  $P@k$  scores for all fine-tuned encoder-based models in the MULTI-TASK setting for  $k \in \{5, 10, 25, 50, 100\}$ . Results are averages across  $k = 5$  test splits and include standard deviations – indicated with shaded areas around the lines.

classify check-worthiness independently. This is a particularly important limitation, as it can potentially lead to fact verification efforts being wasted on content that is not factual. In contrast, all models except LLaMAntino-3-ANITA-8B rarely assign the opposite combination, +FV -CW, which is instead valid within our framework and represents a consistent portion of annotated posts (27.9%). LLaMAntino-3-ANITA-8B favors either two negative labels (-FV -CW) or two positive labels (+FV +CW), while assigning mixed label combinations significantly less often. A side effect of this is that it produces the -FV +CW combination less frequently than the other models. Overall, our analysis shows that models *i*) tend to avoid the combination +FV -CW, preferring to align the two labels rather than diversifying them, especially when they rely on positive factuality/verifiability, and *ii*) tend to produce the invalid label combination -FV +CW. We stress that this tendency is not due to the examples given in the

<sup>19</sup>In this analysis, we report  $P@k$  scores for encoder-based models only since with decoder-based models it is not possible to get confidence scores for labels generated as part of raw outputs.





**Figure 4:** Percentages of factuality/verifiability (FV) and check-worthiness (CW) label combinations in decoder-based models’ outputs in the SEQ setting, averaged on both prompt languages, plus label combinations in WORTHIT’s gold annotations. -FV +CW is emphasized as problematic (red bars w/ vertical lines), as non-factual/verifiable posts cannot be check-worthy.

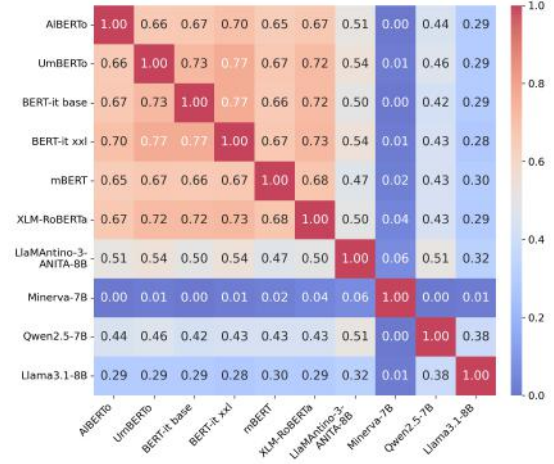
prompts (cf. Table 3), but is rather a general preference of those models, which seem to ignore the relation between factuality/verifiability and check-worthiness.

**Correlation between models’ outputs** To assess if there is a pairwise correlation between encoder- and decoder-based models’ outputs, we calculate the Pearson correlation coefficient ( $r$ ) between all models’ predictions. The heatmap in Figure 5 summarizes the results across the  $k = 5$  test splits. We consider the best-performing setup for each model, namely the MULTI-TASK setting for encoder-based models (see Table 5) and the setup that led to the best performance for each decoder-based model (i.e., language and setting; see Table 6). Encoder-based models exhibit strong positive mutual correlation ( $r \geq 0.65$ ; top-left section in Figure 5), indicating high consistency in the predictions. In contrast, decoder-based models display low inter-model correlation indicating greater output variability. Among them, LlaMAntino-3-ANITA-8B shows the highest alignment with encoder-based models, reaching  $r = 0.54$  with UmBERTo and BERT-it xxl. Conversely, Minerva-7B consistently shows no or very weak correlation with other models – with  $r$  ranging from 0.00 to 0.06 – revealing that its outputs are largely unrelated with those of all other models.

## 6. Conclusion

We introduce WORTHIT, the first dataset of Italian social media posts annotated for factuality/verifiability and check-worthiness that spans multiple years and topics and includes human label variation. We conduct thorough check-worthiness estimation experiments with encoder- and decoder-based models. Results show that the former models in a multi-task setting reach the best results, while the latter models systematically classify non-factual/verifiable posts as check-worthy, failing to capture the relation between the two concepts.

WORTHIT’s partial overlap with a dataset for fallacy detection, FAINA [30], opens new research avenues for



**Figure 5:** Pearson correlation coefficient ( $r$ ) between models’ predictions over the  $k = 5$  test splits, considering the best-performing setup for all encoder- and decoder-based models.

combining the two tasks. Further opportunities include modeling human label variation for the check-worthiness task using the released parallel annotations and experimenting with additional models, training setups, and prompting strategies. Finally, the wide temporal coverage and the diverse set of topics represented in WORTHIT open the field to studies on out-of-distribution generalization of check-worthiness estimation models.

## Acknowledgments

This work has been funded by the European Union’s Horizon Europe research and innovation program under grant agreement No. 101070190 (AI4Trust). We also gratefully acknowledge funding from the German Federal Ministry of Research, Technology and Space (BMFTR) under the grant 01IS23072 for the Software Campus project MULTIVIEW.

## References

- [1] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, *Transactions of the Association for Computational Linguistics* 10 (2022) 178–206. doi:10.1162/tac1\_a\_00454.
- [2] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. S. Kartal, Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets, in: *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, CEUR-WS.org, Bologna, Italy, 2022. URL: <https://ceur-ws.org/Vol-3180/paper-28.pdf>.
- [3] R. Panchendrarajan, A. Zubiaga, Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research, *Natural Language Processing Journal* 7 (2024) 100066. doi:10.1016/j.nlp.2024.100066.
- [4] L. Konstantinovskiy, O. Price, M. Babakar, A. Zubiaga, Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection, *Digital Threats* 2 (2021). doi:10.1145/3412869.
- [5] A. Das, H. Liu, V. Kovatchev, M. Lease, The state of human-centered NLP technology for fact-checking, *Information Processing & Management* 60 (2023) 103219. doi:10.1016/j.ipm.2022.103219.
- [6] P. Atanasova, L. Márquez, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, W. Zaghouani, S. Kyuchukov, G. Da San Martino, P. Nakov, Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. Task 1: Check-worthiness, in: *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR-WS.org, Avignon, France, 2018. URL: [https://ceur-ws.org/Vol-2125/invited\\_paper\\_13.pdf](https://ceur-ws.org/Vol-2125/invited_paper_13.pdf).
- [7] A. Ramponi, B. Plank, Neural unsupervised domain adaptation in NLP—A survey, in: D. Scott, N. Bel, C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 6838–6855. URL: <https://aclanthology.org/2020.coling-main.603/>. doi:10.18653/v1/2020.coling-main.603.
- [8] B. Plank, The “problem” of human label variation: On ground truth in data, modeling and evaluation, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 10671–10682. URL: <https://aclanthology.org/2022.emnlp-main.731/>. doi:10.18653/v1/2022.emnlp-main.731.
- [9] M. Poesio, R. Artstein, The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account, in: A. Meyers (Ed.), *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 76–83. URL: <https://aclanthology.org/W05-0311/>.
- [10] L. Aroyo, C. Welty, Truth is a lie: Crowd truth and the seven myths of human annotation, *AI Magazine* 36 (2015) 15–24. doi:10.1609/aimag.v36i1.2564.
- [11] F. Cabitza, A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, *Proceedings of the AAAI Conference on Artificial Intelligence* 37 (2023) 6860–6868. doi:10.1609/aaai.v37i6.25840.
- [12] Y. Nie, X. Zhou, M. Bansal, What can we learn from collective human opinions on natural language inference data?, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, 2020, pp. 9131–9143. URL: <https://aclanthology.org/2020.emnlp-main.734/>. doi:10.18653/v1/2020.emnlp-main.734.
- [13] P. Atanasova, P. Nakov, G. Karadzhov, M. Moughtarami, G. Da San Martino, Overview of the CLEF-2019 CheckThat! lab: Automatic identification and verification of claims. Task 1: Check-worthiness, in: *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, CEUR-WS.org, Lugano, Switzerland, 2019. URL: [https://ceur-ws.org/Vol-2380/paper\\_269.pdf](https://ceur-ws.org/Vol-2380/paper_269.pdf).
- [14] 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, in: *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, CEUR-WS.org, Thessaloniki, Greece, 2020. URL: [https://ceur-ws.org/Vol-2696/paper\\_265.pdf](https://ceur-ws.org/Vol-2696/paper_265.pdf).
- [15] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, A. Nikolov, M. Kutlu, Y. S. Kartal, F. Alam, G. Da San Martino, A. Barrón-Cedeño, R. Míguez, J. Beltrán, T. Elsayed, P. Nakov, Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates, in: *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*, CEUR-WS.org, Bucharest, Romania, 2021. URL: <https://ceur-ws.org/Vol-2936/paper-28.pdf>.
- [16] F. Alam, A. Barrón-Cedeño, G. S. Cheema, G. K.

- Shahi, S. Hakimov, M. Hasanain, C. Li, R. Míguez, H. Mubarak, W. Zaghouani, P. Nakov, Overview of the CLEF-2023 CheckThat! lab task 1 on check-worthiness in multimodal and multigenre content, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), CEUR-WS.org, Thessaloniki, Greece, 2023. URL: <https://ceur-ws.org/Vol-3497/paper-019.pdf>.
- [17] M. Hasanain, R. Suwaileh, S. Weering, C. Li, T. Caselli, W. Zaghouani, A. Barrón-Cedeño, P. Nakov, F. Alam, Overview of the CLEF-2024 CheckThat! lab task 1 on check-worthiness estimation of multigenre content, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), CEUR-WS.org, Grenoble, France, 2024. URL: <https://ceur-ws.org/Vol-3740/paper-24.pdf>.
- [18] N. Salek Faramarzi, F. Hashemi Chaleshtori, H. Shirazi, I. Ray, R. Banerjee, Claim extraction and dynamic stance detection in COVID-19 tweets, in: Companion Proceedings of the ACM Web Conference 2023, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1059–1068. doi:10.1145/3543873.3587643.
- [19] R. Dhar, D. Das, Leveraging expectation maximization for identifying claims in low resource Indian languages, in: S. Bandyopadhyay, S. L. Devi, P. Bhattacharyya (Eds.), Proceedings of the 18th International Conference on Natural Language Processing (ICON), NLP Association of India (NLPAI), Silchar, India, 2021, pp. 307–312. URL: <https://aclanthology.org/2021.icon-main.37>.
- [20] J. Gili, L. Passaro, T. Caselli, Check-IT!: A corpus of expert fact-checked claims for Italian, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023), CEUR Workshop Proceedings, Venice, Italy, 2023, pp. 227–235. URL: <https://aclanthology.org/2023.clicit-1.29/>.
- [21] A. Scaiella, S. Costanzo, E. Passone, D. Croce, G. Gambosi, Leveraging large language models for fact verification in Italian, in: F. Dell’Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 898–908. URL: <https://aclanthology.org/2024.clicit-1.97/>.
- [22] P. Atanasova, D. Wright, I. Augenstein, Generating label cohesive and well-formed adversarial claims, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 3168–3177. URL: <https://aclanthology.org/2020.emnlp-main.256/>. doi:10.18653/v1/2020.emnlp-main.256.
- [23] G. Valer, A. Ramponi, S. Tonelli, When you doubt, abstain: A study of automated fact-checking in Italian under domain shift, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023), CEUR Workshop Proceedings, Venice, Italy, 2023, pp. 433–440. URL: <https://aclanthology.org/2023.clicit-1.52/>.
- [24] E. M. Williams, P. Rodrigues, S. Tran, Accenture at CheckThat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, Bucharest, Romania, 2021, pp. 659–669. URL: <https://ceur-ws.org/Vol-2936/paper-55.pdf>.
- [25] R. A. Frick, I. Vogel, J.-E. Choi, Fraunhofer SIT at CheckThat! 2023: Enhancing the detection of multimodal and multigenre check-worthiness using optical character recognition and model souping, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), CEUR-WS.org, Thessaloniki, Greece, 2023, pp. 337–350. URL: <https://ceur-ws.org/Vol-3497/paper-029.pdf>.
- [26] M. Sawinski, K. Wecel, E. Ksiezniak, M. Stróżyńska, W. Lewoniewski, P. Stolarski, W. Abramowicz, OpenFact at CheckThat! 2023: Head-to-head GPT vs. BERT - A comparative study of transformers language models for the detection of check-worthy claims, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), CEUR-WS.org, Thessaloniki, Greece, 2023, pp. 453–472. URL: <https://ceur-ws.org/Vol-3497/paper-040.pdf>.
- [27] A. Savchev, AI Rational at CheckThat! 2022: Using transformer models for tweet classification, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, Bologna, Italy, 2022, pp. 656–659. URL: <https://ceur-ws.org/Vol-3180/paper-52.pdf>.
- [28] Y. Li, R. Panchendrarajan, A. Zubiaga, FactFinders at CheckThat! 2024: Refining check-worthy statement detection with LLMs through data pruning, in: G. Faggioli, N. Ferro, P. Galuscáková, A. García Seco de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), CEUR-WS.org, Grenoble, France, 2024, pp. 520–537. URL: <https://ceur-ws.org/Vol-3740/paper-47.pdf>.
- [29] A. F. Hayes, K. Krippendorff, Answering the call for a standard reliability measure for coding data,

- Communication Methods and Measures 1 (2007) 77–89. doi:10.1080/19312450709336664.
- [30] A. Ramponi, A. Daffara, S. Tonelli, Fine-grained fallacy detection with human label variation, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 762–784. URL: <https://aclanthology.org/2025.naacl-long.34/>. doi:10.18653/v1/2025.naacl-long.34.
- [31] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, ALBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets, in: R. Bernardi, R. Navigli, G. Semeraro (Eds.), *Proceedings of the Sixth Italian Conference on Computational Linguistics, CEUR-WS.org, Bari, Italy, 2019*. URL: <https://ceur-ws.org/Vol-2481/paper57.pdf>.
- [32] L. Parisi, S. Francia, P. Magnani, UmbERTO: An Italian language model trained with whole word masking, 2020. URL: <https://github.com/musixmatchresearch/umberto>, accessed: 2025-05-01.
- [33] S. Schweter, Italian BERT and ELECTRA models, 2020. doi:10.5281/zenodo.4263142, accessed: 2025-05-01.
- [34] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423/>. doi:10.18653/v1/N19-1423.
- [35] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020*, pp. 8440–8451. URL: <https://aclanthology.org/2020.acl-main.747/>. doi:10.18653/v1/2020.acl-main.747.
- [36] R. van der Goot, A. Üstün, A. Ramponi, I. Sharaf, B. Plank, Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP, in: D. Gkatzia, D. Seddah (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2021*, pp. 176–197. URL: <https://aclanthology.org/2021.eacl-demos.22/>. doi:10.18653/v1/2021.eacl-demos.22.
- [37] M. Polignano, P. Basile, G. Semeraro, LLaMAntino-3-ANITA-8B-Inst-DPO-ITA model, 2024. URL: <https://huggingface.co/swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA>, accessed: 2025-05-01.
- [38] R. Orlando, L. Moroni, P.-L. Huguet Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, Minerva LLMs: The first family of large language models trained from scratch on Italian data, in: F. Dell’Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024*, pp. 707–719. URL: <https://aclanthology.org/2024.clicit-1.77/>.
- [39] Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, et al., Qwen2.5 technical report, arXiv preprint arXiv:2412.15115 (2025). URL: <https://arxiv.org/abs/2412.15115>.
- [40] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, et al., The Llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024). URL: <https://arxiv.org/abs/2407.21783>.



## Appendix

### A. Search Keywords

We report the full list of search keywords, divided by topic, in Table 7. Within squared brackets are the grammatical gender and number variants (if any) that we included for each keyword.

### B. Annotation Guidelines

For factuality/verifiability annotation, a post can be either factual/verifiable (i.e., YES label) or non factual/verifiable (i.e., NO). For posts that are factual/verifiable, a check-worthiness label in a 5-point Likert scale must also be assigned. Possible labels are: DEFINITELY YES, PROBABLY YES, NEITHER YES NOR NO, PROBABLY NO, and DEFINITELY NO. For both annotation tasks, we strictly follow the guidelines by Nakov et al. [2] and translate them to Italian. The annotation guidelines are presented below.

#### Factuality/verifiability

*Il post contiene un'affermazione fattuale che può essere verificata? A titolo di esempio, sono fattuali/verificabili i post che riportano una definizione, menzionano una quantità nel presente o nel passato, fanno una previsione verificabile del futuro, fanno riferimento a leggi, procedure e norme operative, discutono di immagini o video, e indicano correlazioni o causalità.*


#### Check-worthiness


*Credi che l'affermazione contenuta nel post dovrebbe essere verificata da un fact-checker professionista? Questa domanda richiede un giudizio soggettivo basato sulle seguenti domande:*


- 1. L'affermazione espressa nel post potrebbe essere falsa?*
- 2. L'affermazione espressa nel post potrebbe essere di interesse pubblico e/o avere impatto sulla collettività?*
- 3. L'affermazione espressa nel post potrebbe danneggiare la società, un gruppo, un singolo o un'entità?*


*L'annotazione è necessaria solo se il post è stato classificato come fattuale/verificabile. Nota: affermazioni facilmente verificabili dagli utenti (es. "Gli abitanti della Cina sono la metà di quelli dell'Italia") non sono da ritenere check-worthy.*


In the guidelines, we further include information on how to deal with special cases to minimize ambiguity. All the cases provided to annotators are outlined below.


 **Reported speech**, including quotations, references to newspaper and TV, is always factual/verifiable. E.g.: “‘What is done to migrants is criminal’ #PopeFrancis on #CTCF #Rai3” is **factual/verifiable**


 If the claim is in a **subordinate clause**, the post is not factual/verifiable. However, it is factual/verifiable if the claim is salient and conveys the main information. E.g.: “Dear #novax who appeals to art.32 of the Constitution, you should know that the Constitutional Court with ruling no. 307/1990 has decided that a treatment can become mandatory if it serves to protect oneself and the health of others. So, if needed, you vaccinate or leave.” is **factual/verifiable**


 **Generic sentences** are not factual/verifiable because they contain imprecise information (e.g., frequent use of indefinite quantifiers such as *various, some, many*). E.g.: “Three months after the collapse of the #MorandiBridge. From the government only many promises, zero facts and a totally insufficient decree.” is **not factual/verifiable**

 **Personal opinions** are not factual/verifiable, as there is no clear evidence to support them. E.g.: “Put Salvini back at the Interior Ministry, he is the only one who can handle migrants arrivals.” is **not factual/verifiable**

 When the **implicit subject** can be reconstructed, the sentence can be factual/verifiable. E.g.: “When he was minister and closed the ports he said go ahead and prosecute me. Then he was investigated and hid behind parliamentary immunity. When he was minister he insulted Carola Rackete. Then they propose him a TV debate with her and he declines the invitation. And they call him Captain.” is **factual/verifiable**

 **Descriptions of images/videos** with URLs are factual/verifiable when they contain an externally verifiable fact. E.g.: “I receive directly from a Sudanese boy these images. The migrants are leaving the UNHCR center 15 km from #Agadez and marching towards the city.” is **factual/verifiable**

 Posts about **weather conditions** or **temperatures** are considered factual/verifiable when the information is precise, they specify the type of event described, the exact location and time. Posts about temperature are not check-worthy. E.g.: “The situation now in #Catania. I think there is a small problem with climate change. [URL]” is **not factual/verifiable**

 Posts describing **events** (demonstrations, marches, strikes, rallies, initiatives, assemblies, meetings, presentations) are always factual/verifiable. They can include the expressions *everyone for, see you on, together with*. They are generally not check-worthy. E.g.: “#StopFalsePromises! In the streets of Rome with [USER] for global climate strike! #ClimateStrike” is **factual/verifiable**

**Table 7**

Search keywords used for collecting posts in WORTHIT, with grammatical gender and number variants (if any) indicated using squared brackets. Note that these exactly match the keywords that have been used to collect the FAINA dataset [30].

<b>MIGRATION:</b> apolid[e,i]; apolidia; centr[o,i] di accoglienza; centr[o,i] di identificazione ed espulsione; centr[o,i] di permanenza per il rimpatrio; centri di permanenza per i rimpatri; centr[o,i] di permanenza temporanea; centr[o,i] per il rimpatrio; centri per i rimpatri; corridio[i,o,i] umanitar[i,o,i]; domand[a,e] d'asilo; domand[a,e] di asilo; emigrant[e,i]; emigrat[o,i,a,e]; emigrazion[e,i]; espatr[i,o,i]; fattor[e,i] di spinta; immigrant[e,i]; immigrat[o,i,a,e]; immigrazione[i,e,i]; ius sanguinis; migrant[e,i]; migrator[i,o,i,i,a,e]; migrazion[e,i]; minor[e,i] stranieri[o,i] non accompagnat[o,i]; minor[e,i] stranieri[a,e] non accompagnat[a,e]; non-refoulemen[t,ts]; permess[o,i] di soggiorno; procedur[a,e] d'asilo; procedur[a,e] di asilo; protezion[e,i] sussidiari[a,e]; protezion[e,i] umanitari[a,e]; push facto[r,rs]; refoulemen[t,ts]; reinsediament[o,i]; respingiment[o,i]; richiedent[e,i] asilo; rifugiat[o,i,a,e]; rimpatri[o,i]; rimpatriat[o,i,a,e]; sfollat[o,i,a,e]; vittim[a,e] della tratta; vittim[a,e] di tratta
<b>CLIMATE CHANGE:</b> acidificazione dell'oceano; acidificazione degli oceani; aerosol atmosferic[o,i]; allagament[o,i]; alluvion[e,i]; alluvional[e,i]; ambientalismo di facciata; anidride carbonica; antropocene; aridità; bilanc[i,o,i] climatic[o,i]; bilanc[i,o,i] energetic[o,i]; bilanc[i,o,i] idrologic[o,i]; biocombustibil[e,i]; biodegradabil[e,i]; biodegradabilità; biodiversità; biossido di carbonio; cambiament[o,i] climatic[o,i]; cambiament[o,i] del clima; carbon cost; carbon footprint; carbon pricing; carbon tax; cost[o,i] del carbonio; climate; climate change; climate crisis[is,es]; climatic[o,a,i,he]; climatologia; co2; combustibil[e,i] fossil[e,i]; confin[e,i] planetar[i,o,i]; consum[o,i] di suolo; crisi climatic[a,he]; deforestazione[i,e,i]; desalinizzazione[i,e,i]; desertificazione[i,e,i]; diossido di carbonio; disboscament[o,i]; dissalazion[e,i]; ecological footprint; ecologismo di facciata; economi[a,e] circolar[e,i]; effetto serra; emission[e,i]; energi[a,e] rinnovabil[e,i]; esondazion[e,i]; event[o,i] meteorologic[o,i] estrem[o,i]; fenomen[o,i] meteorologic[o,i] estrem[o,i]; finanza sostenibile; fonte di energia rinnovabile; fonti di energia rinnovabil[e,i]; forzant[e,i] radiativ[o,i]; gas serra; gas silvestre; glacialis[m,o,i]; glaciazion[e,i]; greenwashing; impronta carbonica; impronta di carbonio; impronta ecologica; innalzamento de[l,i] mar[e,i]; innalzamento del livello de[l,i] mar[e,i]; innalzamento dei livelli de[l,i] mar[e,i]; inondazion[e,i]; inquinamento atmosferico; inquinamento dell'atmosfera; isol[a,e] di calore; isol[a,e] urban[a,e] di calore; limit[e,i] planetar[i,o,i]; meteorologia; microclima; mobilità sostenibile; mutament[o,i] climatic[o,i]; olocene; ondat[a,e] di caldo; ondat[a,e] di calore; paleoclima; particolato; pedoclima; permafrost; permagelo; prezz[o,i] del carbonio; proiezion[e,i] climatic[a,he]; report di sostenibilità; riscaldamento climatico; riscaldamento globale; risch[i,o,i] climatic[o,i]; scenar[i,o,i] climatic[o,i]; sciogliment[o,i] dei ghiacciai; siccità; sistem[a,i] climatic[o,i]; sostenibilità ambientale; surriscaldamento climatico; surriscaldamento globale; svilupp[o,i] sostenibil[e,i]; tass[a,e] sul carbonio; transizion[e,i] ecologic[a,he]; transizion[e,i] energetic[a,he]; uso d[e,i] suolo; utilizzazione[e,i] del suolo; utilizzo d[e,i] suolo; variabilità climatic[a,he]
<b>PUBLIC HEALTH:</b> agend[a,e] di prenotazione; alfabetizzazione alla salute; alfabetizzazione sanitaria; assistenz[a,e] domiciliar[e,i]; assistenz[a,e] ospedaliere[a,e]; assistenz[a,e] sanitari[a,e]; assistenza universale; aziend[a,e] ospedaliere[a,e]; aziend[a,e] sanitari[a,e]; bisogn[o,i] sanitar[i,o,i]; calendar[i,o,i] di prenotazione; caric[o,hi] di malattia; centro unificato di prenotazione; città san[a,e]; class[e,i] di priorità; comportament[o,i] a rischio; comportament[o,i] di salute; copertur[a,e] sanitari[a,e]; copertur[a,e] universal[e,i]; cur[a,e] medic[a,he]; cur[a,e] sanitari[a,e]; degent[e,i]; degenz[a,e]; determinant[e,i] della salute; determinant[e,i] di salute; dimission[e,i] ospedaliere[a,e]; dispositiv[o,i] medic[o,i]; disuguaglianz[a,e] di salute; disuguaglianz[a,e] nella salute; disuguaglianz[a,e] sanitari[a,e]; educazione alla salute; educazione sanitaria; epidem[i,a,e]; epidemic[o,a,i,he]; epidemiologia; epidemiologic[o,a,i,he]; equità di salute; equità nella salute; equità sanitari[a,e]; esenzion[e,i] dal ticket; esenzion[e,i] ticket; fattor[e,i] di rischio; indicator[e,i] di salute; investment[o,i] nella sanità; investment[o,i] per la salute; investment[o,i] per la sanità; isol[a,e] san[a,e]; istitut[o,i] di cura; istituto di sanità pubblica; istituto superiore di sanità; list[a,e] di attesa; malatti[a,e] infettiv[a,e]; ministero della salute; ministero della sanità; misur[a,e] sanitari[a,e]; ospedali; ospedaliere[o,i,a,e]; ospedalizzazione[e,i]; ospitalizzazione[e,i]; pandem[i,a,e]; politic[a,he] sanitari[a,e]; post[o,i] letto; prestazion[e,i] ambulatorial[e,i]; prestazion[e,i] sanitari[a,e]; prestazion[e,i] specialistic[a,he] ambulatorial[e,i]; prevenzione delle malattie; prevenzione di malattie; prevenzione primaria; prevenzione sanitaria; prevenzione secondaria; prevenzione terziaria; programmazione[e,i] sanitari[a,e]; promozione della salute; promozione di salute; pronto soccorso; ricover[o,i]; salute globale; salute per tutti; salute pubblica; sanità; sanità pubblica; sanitar[i,o,i,i,a,e]; serviz[i,o,i] infermieristic[o,i]; serviz[i,o,i] medic[o,i]; serviz[i,o,i] sanitar[i,o,i]; settore[e,i] sanitar[i,o,i]; sicurezza dell[a,e] cur[a,e]; struttur[a,e] di ricovero; struttur[a,e] ospedaliere[a,e]; struttur[a,e] sanitari[a,e]; terapi[a,e] intensiv[a,e]; trattament[o,i] di salute; trattament[o,i] medic[o,i]; trattament[o,i] sanitar[i,o,i]; uguaglianz[a,e] di salute; uguaglianz[a,e] nella salute; uguaglianz[a,e] sanitari[a,e]; vaccin[o,i]; vaccinazione[e,i]

## C. Hyperparameters

For encoder-based models, we use default MaChAmp (v0.4.2) [36] hyperparameter values and tune the most crucial ones during development. The search space for them is indicated within brackets in Table 8, with best values underlined. The best loss weight value for the auxiliary factuality/verifiability task is set to 0.50 for UmBERTo, BERT-it base, and mBERT, to 0.75 for XLM-RoBERTa, and to 1.00 for ALBERTo and BERT-it xxl.

For decoder-based models, we use the Hugging Face Transformers library using default hyperparameter values and setting the max\_new\_tokens parameter to 30. Since all models are instruction-tuned, we structure our inputs as conversational prompts using the following format: {"role": "user", "content": "prompt"}.

## D. Prompts and Examples

We present the prompt templates used for factuality/verifiability and check-worthiness tasks. For prompts using guidelines, \$[FV|CW]\_GUIDELINES placeholders are replaced with text in the desired language from Table 9. \$[FV|CW]\_EXAMPLES placeholders are replaced with

**Table 8**

Hyper-parameter values employed for encoder-based models.

Hyperparameter	Value
Optimizer, $\beta_1, \beta_2$	AdamW, 0.9, 0.99
Dropout	0.2
Epochs	3
Batch size	{32, 64}
Learning rate	{1e-4, 1e-5}
LR scheduler	Slanted triangular
Weight decay	0.01
Decay factor, cut fraction	0.38, 0.3
Class weights	{balanced, unbalanced}
Main task loss weight ( $\lambda_{cw}$ )	1.00
Aux task loss weight ( $\lambda_{fv}$ )	{0.25, <u>0.50</u> , <u>0.75</u> , <u>1.00</u> }

text-label pair examples in the format “\$POST\_TEXT = \$POST\_LABEL”, one per line. We report the final example set in Table 10. Finally, \$POST\_TEXT is replaced with the text of the post to classify. The first part of the check-worthiness prompt (i.e., en: “You ... Now” and it: “Hai ... Ora”) is included only in the SEQ setting, with \$FV\_LABEL representing the factuality/verifiability label obtained for the same post using the factuality/verifiability prompt.

Table 9

Guidelines for both tasks in Italian and English used for prompting decoder-based models in configurations with guidelines.

<p><b>\$FV_GUIDELINES Italian:</b> "Linee guida:\Un post è fattuale quando contiene informazioni salienti che possono essere verificate esternamente. Tali informazioni possono essere trovate ovunque, comprese subordinate, sostantivi e hashtag. I discorsi riportati e le citazioni sono sempre fattuali. Anche i post che descrivono eventi e attività sono sempre fattuali. I post sul meteo o sulla temperatura e le descrizioni di foto e video sono fattuali solo quando le informazioni sono precise e la località è nota. Al contrario, le affermazioni generiche o vaghe e le opinioni personali non sono fattuali perché non esistono prove chiare a sostegno." <b>English:</b> "Guidelines:\A post is factual when it contains salient information that can be externally verified. Such information can be found everywhere, including subordinates clauses, nouns and hashtags. Reported discourses and references are always factual. Similarly, posts describing events and activities are always factual. Posts about weather or temperature, as well as photo and video descriptions, are factual only when the information is precise and the location is known. On the other hand, generic or vague statements and personal opinions are not factual because there is no clear evidence to support them."</p>
<p><b>\$CW_GUIDELINES Italian:</b> "Linee guida:\Un post può essere check-worthy solo se è fattuale. Un post è considerato check-worthy se è rilevante per la società e può causare danno o modificare le opinioni delle persone. Le affermazioni generiche e le opinioni non sono check-worthy. I post che descrivono eventi climatici e meteorologici di solito non sono check-worthy perché non contengono informazioni sensibili. Allo stesso modo, i post che menzionano che una specifica attività è in corso di svolgimento di solito non sono check-worthy." <b>English:</b> "Guidelines:\A post can be check-worthy only if it is factual. A post is check-worthy if it is relevant to society and can cause harm or modify people's opinions. Generic statements and opinions are not check-worthy. Posts describing climate and weather events are usually not check-worthy because they do not contain sensitive information. Similarly, posts mentioning that a specific activity is taking place are usually not check-worthy."</p>

Table 10

Examples used for few-shot decoder-based models' prompting on the test set. Examples refer to set #1 (see Table 2).

Post text	FV	CW
<p>è solo maggio. e questo #caldo mi terrorizza. ecco. l'ho detto. #crisisclimatica cosa diamine stiamo aspettando???</p> <p><i>it's only May. and this #heat terrifies me. there. I said it. #climatecrisis what the hell are we waiting for???</i></p>	-	-
<p>ma è tipo la seconda volta che i rifugiati recuperati in mare sono 49. mi è preso il sospetto che la libia stia trollando salvini.</p> <p><i>but it's like the second time that the refugees rescued at sea are 49. I got the suspicion that Libya is trolling Salvini.</i></p>	+	+
<p>ho scritto e riscritto che #inceneritore è proposta anti-europea: ue avrebbe eliminato esenzione dell'incenerimento dal pagamento co2 non più tardi del 2028 perché dannoso e rendendolo ancora meno conveniente. sono stato smentito: oggi hanno votato. dal 2026! [URL]</p> <p><i>I've written and rewritten that the #incinerator is an anti-European proposal: the EU would have removed the exemption of incineration from CO2 payments no later than 2028 because it's harmful, making it even less cost-effective. I was contradicted: they voted today. from 2026! [URL]</i></p>	+	+
<p>il fatto che zaia rivoglia il personale "novax" sospeso è la certificazione del danno procurato alla salute pubblica per scelte politiche scellerate e criminali. semplice.</p> <p><i>the fact that Zaia wants the suspended "novax" staff back is proof of the damage caused to public health by reckless and criminal political decisions. simple.</i></p>	+	+
<p>lei pensa ai fratelli migranti in serbia [URL]</p> <p><i>she thinks of the migrant brothers in Serbia [URL]</i></p>	-	-

#### Prompt for factuality/verifiability (en)

Classify the post as "factual" or "not factual".  
 Answer only with "factual" or "not factual".  
 \$FV\_GUIDELINES  
 Examples:  
 \$FV\_EXAMPLES  
 Answer:  
 \$POST\_TEXT =

#### Prompt for factuality/verifiability (it)

Classifica il post come "fattuale" o "non fattuale".  
 Rispondi solo con "fattuale" o "non fattuale".  
 \$FV\_GUIDELINES  
 Esempi:  
 \$FV\_EXAMPLES  
 Risposta:  
 \$POST\_TEXT =

#### Prompt for check-worthiness (en)

You classified the post as \$FV\_LABEL. Now classify the post as "check-worthy" or "not check-worthy". Answer only with "check-worthy" or "not check-worthy".  
 \$CW\_GUIDELINES  
 Examples:  
 \$CW\_EXAMPLES  
 Answer:  
 \$POST\_TEXT =

#### Prompt for check-worthiness (it)

Hai classificato il post come \$FV\_LABEL. Ora classifica il post come "check-worthy" o "non check-worthy". Rispondi solo con "check-worthy" o "non check-worthy".  
 \$CW\_GUIDELINES  
 Esempi:  
 \$CW\_EXAMPLES  
 Risposta:  
 \$POST\_TEXT =