

# Overview of the CLEF-2021 CheckThat! Lab Task 1 on Check-Worthiness Estimation in Tweets and Political Debates

Shaden Shaar<sup>1</sup>, Maram Hasanain<sup>2</sup>, Bayan Hamdan<sup>3</sup>, Zien Sheikh Ali<sup>2</sup>, Fatima Haouari<sup>2</sup>, Alex Nikolov<sup>4</sup>, Mucahid Kutlu<sup>5</sup>, Yavuz Selim Kartal<sup>5</sup>, Firoj Alam<sup>1</sup>, Giovanni Da San Martino<sup>6</sup>, Alberto Barrón-Cedeño<sup>7</sup>, Rubén Míguez<sup>8</sup>, Javier Beltrán<sup>8</sup>, Tamer Elsayed<sup>2</sup> and Preslav Nakov<sup>1</sup>

<sup>1</sup>*Qatar Computing Research Institute, HBKU, Doha, Qatar*

<sup>2</sup>*Qatar University, Qatar*

<sup>3</sup>*Independent Researcher*

<sup>4</sup>*Sofia University, Bulgaria*

<sup>5</sup>*TOBB University of Economics and Technology, Turkey*

<sup>6</sup>*University of Padova, Italy*

<sup>7</sup>*DIT, Università di Bologna, Italy*

<sup>8</sup>*Newtral Media Audiovisual, Spain*

## Abstract

We present an overview of Task 1 of the fourth edition of the CheckThat! Lab, part of the 2021 Conference and Labs of the Evaluation Forum (CLEF). The task asks to predict which posts in a Twitter stream are worth fact-checking, focusing on COVID-19 and politics in five languages: Arabic, Bulgarian, English, Spanish, and Turkish. A total of 15 teams participated in this task and most submissions managed to achieve sizable improvements over the baselines using Transformer-based models such as BERT and RoBERTa. Here, we describe the process of data collection and the task setup, including the evaluation measures, and we give a brief overview of the participating systems. We release to the research community all datasets from the lab as well as the evaluation scripts, which should enable further research in check-worthiness estimation for tweets and political debates.

## Keywords

Check-Worthiness Estimation, Fact-Checking, Veracity, Verified Claims Retrieval, Detecting Previously Fact-Checked Claims, Social Media Verification, Computational Journalism, COVID-19


---

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ sshaar@hbku.edu.qa (S. Shaar); maram.hasanain@qu.edu.qa (M. Hasanain); bayan.hamdan995@gmail.com (B. Hamdan); zs1407404@qu.edu.qa (Z. S. Ali); 200159617@qu.edu.qa (F. Haouari); alexnickolow@gmail.com (A. Nikolov); m.kutlu@etu.edu.tr (M. Kutlu); ykartal@etu.edu.tr (Y. S. Kartal); fialam@hbku.edu.qa (F. Alam); dasan@math.unipd.it (G. D. S. Martino); a.barron@unibo.it (A. Barrón-Cedeño); ruben.miguez@newtral.es (R. Míguez); javier.beltran@newtral.es (J. Beltrán); telsayed@qu.edu.qa (T. Elsayed); pnakov@hbku.edu.qa (P. Nakov)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

# 1. Introduction

The spread of fake news, misinformation and disinformation on the web, in social media and in other communication channels has become an urgent social and political issue. There has been growing interest in fighting against such false or misleading information both in academia and in industry. To address the issue, a number of initiatives have been launched to perform manual claim verification, with over 200 fact-checking organizations worldwide,<sup>1</sup> such as PolitiFact, FactCheck, Snopes, Full Fact, and Newtral, among others. Unfortunately, these efforts are insufficient, given the scale of disinformation propagating in various communication channels, which, in the time of COVID-19, has now grown into the First Global Infodemic (according to the World Health Organization). To deal with this problem, we have launched the CheckThat ! Lab, which features a number of tasks aiming to help automate the fact-checking process and to reduce the spread of misinformation and disinformation.

The CheckThat ! lab was run for the fourth time in the framework of CLEF 2021.<sup>2</sup> The purpose of the 2021 edition of the lab was to foster the development of technology that would enable finding check-worthy claims, detecting claims that have been previously fact-checked, and predicting the veracity of news articles and their topics.

Figure 1 shows the full CheckThat ! identification and verification pipeline, including the tasks on detecting previously fact-checked claims and predicting the veracity and the topic of news articles. Here, we focus on Task 1: check-worthiness estimation.<sup>3</sup> This task focuses on *tweets* and *political debates and speeches*. It consists of the following two subtasks:

**Subtask 1A Check-worthiness of tweets.** Given a topic and a stream of potentially related tweets, rank the tweets according to their check-worthiness for the topic.

**Subtask 1B Check-worthiness of debates or speeches.** Given a political debate/speech, return a list of its sentences, ranked by their check-worthiness.

Subtask 1A was offered in Arabic, Bulgarian, English, Spanish, and Turkish. We focused on COVID-19, vaccines, and politics, and we crawled and manually annotated tweets between January 2020 and March 2021. The participants were free to work on any language(s) of their interest, and they could also use multilingual approaches that learn from all datasets. However, the evaluation was per language. This subtask attracted 15 teams, and the most successful approaches used Transformers or a combination of embeddings, manually engineered features, and neural networks. Section 3 offers more details.

Subtask 1B was offered in English, and it used PolitiFact as a source of previously fact-checked claims. The task attracted two teams, and a combination of pre-processing, data augmentation, and Transformer-based models performed the best. Section 4 gives more details.

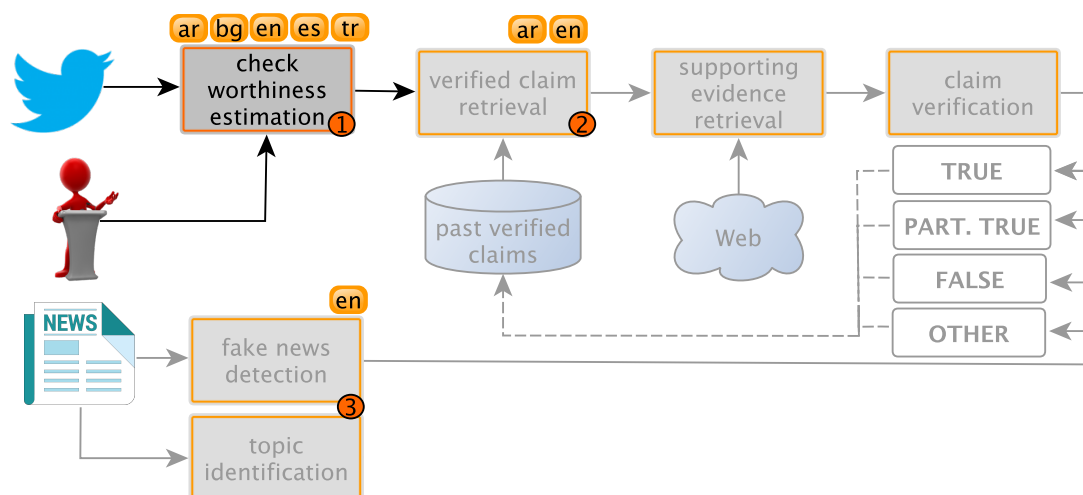
The remainder of the paper is organized as follows: Sections 3 and 4 describe the dataset, the evaluation results, and the participating systems for subtasks 1A and 1B, respectively, Section 2 discusses related work, and Section 5 concludes with final remarks.

---

<sup>1</sup><http://tiny.cc/zd1fnz>; last visited 02/07/2021.

<sup>2</sup><http://sites.google.com/view/clef2021-checkthat/>

<sup>3</sup>Refer to [1] for an overview of the full 2021 edition of the CheckThat ! lab, including the other two tasks on detecting previously fact-checked claims [2] and on fake news detection [3].



**Figure 1:** The full verification pipeline including the three tasks addressed in the CheckThat! lab 2021. Task 1 on check-worthiness estimation focuses on Twitter in five languages (subtask 1A) and debates and speeches in English (subtask 1B). See [2, 3] for a discussion on tasks 2 and 3. The grayed tasks were addressed in previous editions of the lab [4, 5].

## 2. Related Work

Misinformation and disinformation are rapidly spreading in social media, and sometimes false and misleading claims originate in political debates and speeches. To fight the problem, automatic fact-checking has emerged as an important research area, and researchers have worked on a number of subtasks: from automatic identification and verification of claims [6, 7, 8, 9, 5, 10, 11], to identifying check-worthy claims [12, 13, 14, 15], detecting whether a claim has been previously fact-checked [16, 17, 18], retrieving evidence to accept or to reject a claim [19, 20], checking whether the evidence supports or denies the claim [21, 22], and inferring the veracity of the claim [23, 24, 25, 26, 27, 28, 20, 29, 30, 31].

**Check-worthiness estimation for debates/speeches.** The ClaimBuster system [13] was a pioneering work on check-worthiness estimation. Given a sentence in the context of a political debate, it classified it into one of the following, manually annotated categories: *non-factual*, *unimportant factual*, or *check-worthy factual*. In later work, Gencheva & al. [12] also focused on the 2016 US Presidential debates, for which they obtained binary (*check-worthy* vs. *non-check-worthy*) annotations from various fact-checking organizations. An extension of this work resulted in the development of the ClaimRank system, which was trained on more data also including Arabic content [14].

Other related work, also focused on political debates and speeches. For example, Patwari & al. [32] predicted whether a sentence would be selected by a fact-checking organization using a boosting-like model. Similarly, Vasileva & al. [15] used a multi-task learning neural network that predicts whether a sentence would be selected for fact-checking by each individual fact-checking organization (from a set of nine such organizations).

Last but not least, the task was part of the CheckThat! lab in CLEF 2018, 2019, and 2020, where the focus was once again on political debates and speeches, from a single fact-checking organization. In the 2018 edition of the task, a total of seven teams submitted runs for Task 1 (which corresponds to Subtask 1B in 2021), with systems based on word embeddings and RNNs [33, 34, 35, 36]. In the 2019 edition of the task, eleven teams submitted runs for the corresponding Task 1, again using word embeddings and RNNs, and further trying a number of interesting representations [37, 38, 39, 40, 41, 42, 43, 44]. In the 2020 edition of the task, three teams submitted runs for the corresponding Task 5 with systems based on word embeddings and BiLSTM [45], TF.IDF representation with Naïve Bayes, logistic regression, decision trees [46], BERT prediction scores, and word embeddings with logistic regression [47].

**Check-worthiness estimation for tweets.** There has been less effort in identifying check-worthy claims *in social media*. Previous work in this direction includes Task 1 in the 2020 edition of the lab, and the work of Alam et al. [48, 49], who developed a multi-question annotation schema of tweets about COVID-19, organized around seven questions, including one about claim check-worthiness. For some languages in the 2021 Subtask 1A, we used the setup and the annotations for one of the questions in their schema, as well as their data for that question, which we further extend with additional data.

In the 2020 edition of the lab, the focus was on English and Arabic. For the Arabic task, several teams fine-tuned pre-trained models such as AraBERT and multilingual BERT [50, 51, 47]. Other approaches relied on pre-trained models such as GloVe and Word2vec [52, 45] to obtain embeddings for the tweets, which were fed into a neural network or an SVM. In addition to text representations, some teams used other features, namely morphological and syntactic, part-of-speech (POS) tags, named entities, and sentiment features [53, 54]. As for the English task, we also observed the popularity of pre-trained Transformers, namely BERT and RoBERTa [50, 52, 55, 56, 47, 57, 58]. Other approaches relied on word embeddings like GloVe to obtain embeddings for the tweets, which were fed into a neural network [45]. There were also systems that used more traditional machine learning models such as random forest [46].

An indirectly related research line is on credibility assessment of tweets [59], including the CRED BANK tweet corpus [60], which has credibility annotations, multilingual fact-checking corpus [61], as well as work on fake news [62], and on rumor detection in social media [63]; unlike that work, here we focus on detecting check-worthiness rather than predicting the credibility/factuality of the claims in the tweets. Another less relevant research line is on development of datasets of tweets about COVID-19 [64, 65, 66, 67, 68]; however, none of these datasets has focused on check-worthiness estimation.

### 3. Subtask 1A: Check-Worthiness Estimation for Tweets

The aim of Task 1 is to determine whether a piece of text is worth fact-checking. In order to do that, we either resort to the judgments of professional fact-checkers or we ask human annotators to answer several auxiliary questions [49, 48], such as “does it contain a verifiable factual claim?”, “is it harmful?” and “is it of general interest?”, before deciding on the final check-worthiness label.

**Table 1**

**Subtask 1A:** statistics about the CT-CWT-21 dataset for all five languages. The bottom part of the table shows the main topics.

Partition	Arabic	Bulgarian	English	Spanish	Turkish	Total
Training	3,444	3,000	822	2,495	1,899	11,660
Development	661	350	140	1,247	388	2,786
Testing	600	357	350	1,248	1,013	3,568
Total	4,705	3,707	1,312	4,990	3,300	18,014
<b>Main topics</b>						
COVID-19	■	■	■		■	
Politics	■			■	■	

**Subtask 1A** focused on Twitter and it is defined as follows: *“Given a topic and a stream of potentially related tweets, rank the tweets according to their check-worthiness for the topic.”*

The task is offered in Arabic, Bulgarian, English, Spanish, and Turkish. We created and released an independent labeled dataset per language as explained in the following section. The participants were free to work on any language(s) of their interest, and they could also use multilingual approaches that make use of all datasets for training.

### 3.1. Datasets

Although all languages tackled major topics such as COVID-19 and politics, the crawling and the annotation were done differently across the languages due to different resources available to the team leading the annotation for each language. Eventually, for each language, we release a tweet dataset with each tweet labeled for check-worthiness. Below, we provide more detail about how the crawling and the annotation were done for each language. Table 1 shows some statistics about the datasets, which is split into training, development, and testing partitions.

#### 3.1.1. Arabic Dataset

In order to construct the Arabic dataset, we first manually created several topics over the period of several months. Examples of topic titles include “Coronavirus in the Arab World”, “GCC Reconciliation”, and “Deal of the century”. We augmented each topic with a set of keywords, hashtags, and usernames to track in Twitter.<sup>4</sup> Once we had created a topic, we immediately crawled a one-week stream of tweets using the constructed search terms, where we searched Twitter (via the Twitter search API) using each term by the end of each day. We limited the search to original Arabic tweets (i.e., we excluded retweets). We then de-duplicated the tweets and we dropped those matching our qualification filter that excludes tweets containing terms from a blacklist of explicit terms and tweets that contain more than four hashtags or more than two URLs. Afterwards, we ranked the tweets by popularity (defined by the sum of their retweets and likes), and we selected the top-500 to be annotated.

<sup>4</sup>Keywords used to crawl tweets: [http://gitlab.com/checkthat\\_lab/clef2021-checkthat-lab/-/tree/master/task1](http://gitlab.com/checkthat_lab/clef2021-checkthat-lab/-/tree/master/task1)

**Table 2**

**Subtask 1A, Arabic:** tweets with their check-worthiness marked. For each tweet, we also include a translation to English.

✓	الأردن تمنع صينيين من دخول أراضيها بسبب فيروس كورونا. #كورونا_الصيني #الأردن Jordan prevents Chinese citizens from entering the country due to the Corona virus. #Chinese_Corona #Jordan
X	كانت قناة المسيرة ترافق الحوثيين في الجبهات، قبل ان تنضم قناة الجزيرة رسميا للإعلام الحربي الحوثي ضد اليمنيين <a href="https://t.co/bziqkP4CPI">https://t.co/bziqkP4CPI</a> ! Al-Masirah TV was accompanying the Houthis on the war fronts, before Al Jazeera channel officially joined the Houthi war media against the Yemenis! <a href="https://t.co/bziqkP4CPI">https://t.co/bziqkP4CPI</a>
X	نانسي بيلوسي: الرئيس #ترمب يشكل خطرا على الدستور والديمقراطية <a href="https://t.co/pobQlZJs9G">https://t.co/pobQlZJs9G</a> Nancy Pelosi: President Trump is a danger to the constitution and democracy <a href="https://t.co/pobQlZJs9G">https://t.co/pobQlZJs9G</a>
X	عدد أيام الحصار 1310 #المصالحة الخليجية #القمة الخليجية <a href="https://t.co/I25oW4jHBe">https://t.co/I25oW4jHBe</a> 41 The number of siege days is 1310 #gulf_reconciliation #41_gulf_summit <a href="https://t.co/I25oW4jHBe">https://t.co/I25oW4jHBe</a>

The training and the development sets include 12 topics crawled in January, February, and March 2020 and borrowed from the last edition of the CheckThat! lab, considering topics with highest inter-annotation agreement; refer to [51] for further details on the annotation process. For this year's edition, for each topic, we only kept tweets that were relevant and had full inter-rater agreement on the check-worthiness label.

For the test set, we crawled using two topics in January 2021 and we annotated the resulting tweets as follows. We first recruited one annotator to annotate each tweet for its relevance with respect to the target topic. Then, we labeled for check-worthiness the tweets that were found relevant. This second annotation was done by two *expert* annotators. It was followed by a subsequent consolidation step, when the annotators talked to each other to resolve potential disagreements. Due to the subjective nature of check-worthiness, we chose to represent the check-worthiness criteria by several questions, to help the annotators think about different aspects of check-worthiness. First, the annotators were asked to answer the following question:

**1. Does the tweet contain a verifiable factual claim?**

If the answer to the above question is positive, the annotator is asked to answer the following additional yes/no questions:

- 2. Does the claim in the tweet appear to be false?**
- 3. Do you think the claim in the tweet is of interest to or would have an impact on the public?**
- 4. To what extent do you think the claim can morally or physically harm an entity, a country, etc.?**
- 5. Do you think that journalists will be interested in covering the spread of the claim or the information discussed by the claim?**

Once the annotator has answered the above questions, s/he is further required to answer a fifth question considering all the answers given previously:

**6. Do you think that a professional fact-checker or a journalist should verify the claim in the tweet?**

This is a yes/no question and the answer is the label that we will use to represent the check-worthiness for the target tweet. Table 2 shows examples of annotated tweets in Arabic.

### 3.1.2. Bulgarian and English Datasets

We collected tweets that matched the COVID-19 topic using language-specific keywords and hashtags.<sup>5</sup> We ran all the data collection from January 2020 to February 2021, and we selected the most retweeted posts for the manual annotation.

We considered a number of factors for the annotation, including tweet popularity in terms of retweets, which is already taken into account as part of the data collection process. We further asked the annotators to answer the following five questions:<sup>6</sup>

- **Q1: Does the tweet contain a verifiable factual claim?** This is an objective question. Positive examples include tweets that state a definition, mention a quantity in the present or the past, make a verifiable prediction about the future, reference laws, procedures, and rules of operation, discuss images or videos, and state correlation or causation, among others. This is influenced by [69].
- **Q2: To what extent does the tweet appear to contain false information?** This question asks for a subjective judgment; it does not ask for annotating the actual factuality of the claim in the tweet, but rather whether the claim appears to be false.
- **Q3: Will the tweet have an impact on or be of interest to the general public?** This question asks for an objective judgment. Generally, claims that contain information related to potential cures, updates on number of cases, on measures taken by governments, or discussing rumors and spreading conspiracy theories should be of general public interest.
- **Q4: To what extent is the tweet harmful to the society, a person(s), a company(s), or a product(s)?** This question also asks for an objective judgment: to identify tweets that can cause harm.
- **Q5: Do you think that a professional fact-checker should verify the claim in the tweet?** This question asks for a subjective judgment. Yet, its answer should be informed by the answer to questions Q2, Q3 and Q4, as a check-worthy factual claim is probably one that is likely to be false, is of public interest, and/or appears to be harmful. Note that we are stressing the fact that a professional fact-checker should verify the claim, which rules out claims that are easy to fact-check by a layman.

---

<sup>5</sup>The keywords used to crawl the tweets are available at [http://gitlab.com/checkthat\\_lab/clef2021-checkthat-lab/-/tree/master/task1](http://gitlab.com/checkthat_lab/clef2021-checkthat-lab/-/tree/master/task1)

<sup>6</sup>We used the following MicroMappers setup for the annotations:  
<http://micromappers.qcri.org/project/covid19-tweet-labelling/>



We considered as check-worthy the tweets that received a positive answer both to Q1 and to Q5; if there was a negative answer to either Q1 or Q5, the tweet was considered not worth fact-checking. The answers to Q2, Q3, and Q4 were not considered directly, but they helped the annotators make a better decision for Q5. For the task, we did not provide the labels for Q2-Q4.

The annotations were performed by 2–5 annotators independently, and consolidated for the cases of disagreement. The annotation setup was part of a broader initiative; see [48] for details.

**Table 3**

**Subtask 1A, Bulgarian:** tweets with their check-worthiness marked.

<i>I just saw Край на споровете за произхода на COVID-19! Той е изкуствено създаден в САЩ през 2015 г. - Click to see also</i> <a href="https://t.co/ikFXAmp8bO">https://t.co/ikFXAmp8bO</a>	✓
I just saw End of controversy over the origin of COVID-19! It was artificially created in USA in 2015 - Click to see also <a href="https://t.co/ikFXAmp8bO">https://t.co/ikFXAmp8bO</a>	
<i>ЕС установи координация между Русия, Китай и Иран за дезинформиране за COVID-19</i> <a href="https://t.co/84YfO0CA9q">https://t.co/84YfO0CA9q</a>	✓
EU found a coordination between Russia, China and Iran to disinform about COVID-19 <a href="https://t.co/84YfO0CA9q">https://t.co/84YfO0CA9q</a>	
<i>Четвъртия в Руската висша лига Ростов загуби с 10-1 на гости на Сочи, след като трябваше да заложи на младежите до 17 години, защото голяма част от играчите на първия отбор са заразени с Covid-19.</i> <a href="https://t.co/LnA6I0Sx4T">https://t.co/LnA6I0Sx4T</a>	✗
The fourth in the Russian Premier League Rostov lost 10-1 as a guest to Sochi after having to bet on young players under 17 because many of its first team's players were infected with COVID-19. <a href="https://t.co/LnA6I0Sx4T">https://t.co/LnA6I0Sx4T</a>	
<i>(Двама medici от болница "Тракия" са с COVID-19) от Зарата - Новини -</i> <a href="https://t.co/Geg6VEupCJ">https://t.co/Geg6VEupCJ</a> #новини #COVID19 #заразениmedici <a href="https://t.co/vqkH9D2LoU">https://t.co/vqkH9D2LoU</a>	✗
(Two medics from Trakia Hospital have COVID-19) from Zarata - News - <a href="https://t.co/Geg6VEupCJ">https://t.co/Geg6VEupCJ</a> #news # COVID19 #infectedmedics <a href="https://t.co/vqkH9D2LoU">https://t.co/vqkH9D2LoU</a>	

**Table 4**

**Subtask 1A, English:** tweets with their check-worthiness marked.

<i>Breaking: Congress prepares to shutter Capitol Hill for coronavirus, opens telework center</i>	✓
<i>China has 24 times more people than Italy...</i>	✗
<i>Everyone coming out of corona as barista</i>	✗
<i>Lord, please protect my family &amp; the Philippines from the corona virus</i>	✗

Tables 3 and 4 show examples of annotated tweets for Bulgarian and English, respectively. The first English example, ‘*Breaking: Congress prepares to shutter Capitol Hill for coronavirus, opens telework center*’, contains a verifiable factual claim and is of high interest to society, and thus it is check-worthy. The second one, ‘*China has 24 times more people than Italy...*’, contains a verifiable factual claim, but it is trivial to check. The third example, ‘*Everyone coming out of corona as barista*’, is a joke, and thus not check-worthy. The fourth one, ‘*Lord, please protect my family & the Philippines from the corona virus*’, does not contain a verifiable factual claim.



**Table 5**

**Subtask 1A, Spanish:** tweets with their check-worthiness marked (including translations to English).

---

*Un nigeriano de 28 años (aun no sabemos si ilegal o no) es detenido tras intentar raptar a una niña de 11 años en un parque en Getafe. Dicen en la prensa que se desconocen las intenciones del sujeto. Por cierto, Qué montón de hechos aislados! <https://t.co/B4ey4rdg8I>* ✓

A 28 years old Nigerian (we do not know if illegal or not yet) is arrested after trying to kidnap an 11 years old girl in a park at Getafe. The press say that they ignore the intentions of the subject. By the way, how many isolated cases! <https://t.co/B4ey4rdg8I>

*Muy preocupado por el futuro de la plantilla de Endesa en As Pontes. Están en juego los proyectos de 750 familias, el futuro de una comarca y el 50% del tráfico del Puerto de Ferrol. El Gobierno central no puede seguir actuando en todo sin pensar en lo que se lleva por delante* ✗

Quite worried about the future of the staff of Endesa in As Pontes. The projects of 750 families, the future of a region and 50% of the traffic of Puerto de Ferrol are at stake. The central government cannot keep acting without considering the consequences

*70.000.000€ más en guarderías, 65.500.000€ más en escuelas, 129.500.000€ más en universidades, 261 profesores más, 7.365 becas comedor más, un 30% menos en tasas universitarias, 124.000.000€ más en I+D+I, 19.000.000€ más en el Sincrotrón... O somos útiles o no somos.* ✓

70,000,000€ more for nurseries, 65,500,000€ more for schools, 129,500,000€ more for universities, 261 extra professors, 7,365 extra lunch scholarships, 30% less in university tuition fees, 124,000,000€ more for I+D+I, 19,000,000€ more for the Synchrotron... Either we are useful or we are not.

*Una ley muy necesaria, que llevamos defendiendo mucho tiempo, para garantizar el derecho a una muerte digna en nuestro país. <https://t.co/vvpvUd3PDY>* ✗

A much needed law that we have been defending for long time to guarantee the right for a dignified death in our country. <https://t.co/vvpvUd3PDY>

---

### 3.1.3. Spanish Dataset

The dataset consists of 4,990 tweets sampled from the accounts of 350 well-known Spanish politicians. Professional fact-checkers reviewed all tweets published by these accounts over a period of one month to determine whether a tweet contained a check-worthy claim or not. The human fact-checkers considered different editorial criteria to determine whether a tweet is check-worthy, including factuality, public relevance of the character behind the claim, and potential impact on the general audience. Factuality and relevance are the main factors when making the decision. All tweets were annotated independently by three experts and the final decision was made by majority voting.

It is worth noticing that only about 10% of the tweets are considered to be check-worthy, which is close to a realistic distribution. Table 5 shows some examples. We can see that these tweets tend to be fairly long and are often given some context on top of the claim. The third tweet, which is check-worthy, actually contains multiple claims about the impact of a political party on investment. The fourth tweet does contain a claim, but it refers to the support of a new law and was not considered to be check-worthy by the expert.

**Table 6**

**Subtask 1A, Turkish:** tweets with their check-worthiness marked (including translations to English).

*Günün sorusu: 1,5 milyon doz biontech aşısı Türkiye'ye geldi mi? Parasını kim ödedi, ne kadar ve bu aşı kimlere yapıldı?* ✗

The question of the day: Did 1.5M Biontech vaccines arrive Turkey? Who paid it? How much? And who has been vaccinated with them?

*Hükümete vuracağım diye inaktif aşı hakkındaki bilgileri çarpıyorsunuz. Lakin Türkiye'nin kullandığı aşı, diğer aşuların içerisinde hem en güvenilir hem de en etkili Kocan Hocanın eline, emeğine sağlık Süreci manipüle edenleri de Allah nasıl biliyorsa öyle yapsın!* ✓

You are distorting information about inactive vaccines in order to criticize the government. However, the vaccine used in Turkey is the most reliable and the most effective one among other vaccines. Thanks to Kocan Hodga for his efforts. May Allah treat those manipulators as He wants.

Önümüzdeki iki hafta çok kritik Korona ŞehirEfsaneleri ✗

The following two weeks are very critical for Corona UrbanLegends

*45 gün önce aşı olan doktor koronavirüsten hayatını kaybetti <https://t.co/a7qj8I8CMb> <https://t.co/ZiitFoHJmS>* ✓

The doctor who was vaccinated 45 days ago lost his life because of coronavirus <https://t.co/a7qj8I8CMb> <https://t.co/ZiitFoHJmS>

### 3.1.4. Turkish Dataset

For the training and the development sets for Turkish, we used the TrClaim-19 [70] dataset. For the testing dataset, we crawled tweets using keywords related to health and to COVID-19 from February 26, 2021 till March 29, 2021. After de-duplication, we randomly selected tweets for manual annotation. Each tweet was annotated by three annotators, which were asked to say whether the corresponding tweet contains a check-worthy claim or not. The labels were aggregated based on majority voting.

Table 6 shows some examples. The first example does not contain a verifiable factual claim, just questions, and thus it is judged not to be check-worthy. The second tweet contains a factual claim about the effectiveness of the vaccine used in Turkey, which is of wide public interest, and is thus check-worthy. The third claim makes a statement about the future that is not particularly interesting. The last claim is of public interest as it contains a factual claim about a possible mortality case due to a COVID-19 vaccine, and it is thus considered check-worthy.

### 3.2. Evaluation

This is a ranking task, where a tweet has to be ranked according to its check-worthiness. Therefore, we consider **mean average precision (MAP) as the official evaluation measure**, which we complement with reciprocal rank (RR), R-precision (R-P), and P@k for  $k \in \{1, 3, 5, 10, 20, 30\}$ . The data and the evaluation scripts are available online.<sup>7</sup>

<sup>7</sup>[https://gitlab.com/checkthat\\_lab/clef2021-checkthat-lab/](https://gitlab.com/checkthat_lab/clef2021-checkthat-lab/)

Table 7

Subtask 1A: summary of the approaches used by the participating systems.

Team	Models										Other		
	SBERT	BERT	RoBERTa	ALBERT	DistilBERT	Electra	BETO	AraBERT	Ara-ALBERT	BERTurk	Data augmentation	Preprocessing	LIWC
1. abaruah –													
2. Accenture [71]		●	●								●		
3. bigIR –													
4. csum112 –													
5. DamascusTeam –													
6. Fight for 4230 [72]		●									●	●	
7. GPLSI [73]		●					●					●	●
8. iCompass [74]								●	●			●	
9. NLP&IR@UNED [75]		●	●	●	●	●							
10. NLytics [76]			●										
11. QMUL-SDS [77]								●				●	
12. SCUoL [78]								●					
13. SU-NLP [79]										●		●	
14. TOBB ETU [80]		●	●				●	●		●			
15. UPV [81]	●									●			

### 3.3. Overview of the Systems

Fifteen teams took part in this task, with English and Arabic being the most popular languages. Four out of these fifteen teams submitted runs for all five languages —most of them having trained independent models for each language (yet, team UPV trained a single multilingual model). Most of the system were based on state-of-the-art pre-trained Transformers such as BERT [82] and RoBERTa [83]. Table 7 summarizes the approaches used by the primary submissions of the participating teams. We can see that BERT, AraBERT, and RoBERTa were by far the most popular pre-trained language models among the participants.

Below, we provide a short summary of the systems submitted by the participating teams. For each team, we indicate in subscript the languages they took part in and the corresponding rank.

**Team Accenture [71]** (ar: 1 bg: 4 en: 9 es: 5 tr: 5) used pre-trained language models such as BERT and RoBERTa. They further used data augmentation; in particular, they generated synthetic training data using lexical substitution to create additional synthetic examples for the positive class. To find the most probable substitutions, they used BERT-based contextual embeddings. They further added a mean-pooling and a dropout layers on top of the model before the final classification layer.

**Team Fight for 4230 [72]** (en: 2) focused on augmented the data by means of machine translation and WordNet-based substitutions. Pre-processing included link removal and punctuation cleaning, as well as quantity and contraction expansion. All hashtags related to COVID-19 were normalized into one, and hashtags were further expanded. Their best approach was based on BERTweet with a dropout layer and the above-described pre-processing.

**Team GPLSI [73]** (en: 5 es: 2) applied the RoBERTa and the BETO transformers together with different manually engineered features, such as the occurrence of dates and numbers, or words from LIWC. A thorough exploration of parameters was made using weights and biases techniques. They also tried to split the four-class classification into two binary classifications and one three-class classification. Finally, they tried oversampling and undersampling.

**Team iCompass (ar: 4)** used preprocessing, including (i) removing English words, (ii) removing URLs and mentions, and (iii) data normalization by removing tashkeel and the letter madda from texts, as well as duplicates, and replacing some characters to prevent mixing. They proposed a simple ensemble of two BERT-based models, including AraBERT and Arabic-ALBERT.

**Team NLP&IR@UNED [75]** (en: 1 es: 4) used several transformer models, such as BERT, ALBERT, RoBERTa, DistilBERT, and Funnel-Transformer. For their official submissions, for English, they used BERT trained on tweets, while for Spanish, they used Electra.

**Team NLytics [76]** (en: 8) used RoBERTa with a regression function in the final layer by treating the problem as a ranking task.

**Team QMUL-SDS [77]** (ar: 4) used the AraBERT pre-processing (i) to replace URLs, email addresses, and user mentions with standard words, (ii) to remove line breaks, HTML markup, repeated characters, and unwanted characters, e.g., emotion icons, and (iii) to handle white spaces between words and digits (non-Arabic or English), and before/after two brackets; they also (iv) removed unnecessary punctuation. They addressed the task as a ranking problem, and fine-tuned an Arabic transformer (AraBERTv0.2-base) on a combination of the data from this year and the data from the CheckThat! lab 2020 (using the CT20-AR dataset).

**Team SCUoL [78]** (ar: 3) used typical preprocessing steps, and fine-tuned different AraBERT models; eventually, they used AraBERTv2-base.

**Team SU-NLP [79]** (tr: 2) used preprocessing, including (i) removing emojis, hashtags, and (ii) replacing all mentions with a special token (@USER), and all URLs with the website's domain. If the URL was for a tweet, they replaced it with TWITTER and the user account name. Finally, they used an ensemble of BERTurk models fine-tuned using different seed values.

**Team TOBB ETU [80]** (ar: 6 bg: 5 en: 10 es: 1 tr: 1) used data augmentation by machine translation, weak supervision, and cross-lingual training. They removed URLs and user mentions and fine-tuned a separate BERT-based models for each language. In particular, they fine-tuned BERTurk<sup>8</sup>, AraBERT, BETO<sup>9</sup>, and BERT-base for Turkish, Arabic, Spanish, and English, respectively. For Bulgarian, they fine-tuned a RoBERTa model pre-trained on Bulgarian documents.<sup>10</sup>

---

<sup>8</sup><http://huggingface.co/dbmdz/bert-base-turkish-cased>

<sup>9</sup><https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased>

<sup>10</sup><http://huggingface.co/iarfmoose/roberta-base-bulgarian>

**Team UPV [81]** (ar: 8 bg: 2 en: 3 es: 6 tr: 4) used a multilingual sentence transformer (SBERT) with knowledge distillation, originally intended for question answering. They introduced an auxiliary language identification task, aside the downstream check-worthiness task.

### 3.4. Performance for Different Languages

Table 8 shows the performance of the official submissions on the test set, in addition to the  $n$ -gram baseline. The official run was the last valid blind submission by each team. The table shows the runs ranked on the basis of the official MAP measure and includes all five languages.

**Arabic** Eight teams participated for Arabic, submitting a total of 17 runs (yet, recall that only the last submission counts). All participating teams adopted fine-tuning existing pre-trained models, such as AraBERT, and multilingual BERT models. We can see that the top two systems additionally worked on improved training datasets. Team **Accenture** used a label augmentation approach to increase the positive examples, while team **bigIR** augmented the training set with the Turkish training set (which they automatically translated to Arabic).

**Bulgarian** Four teams took part for Bulgarian, submitting a total of 11 runs. The top-ranked team was **bigIR**. They did not submit a task description paper, and thus we cannot give much detail about their system. Team **UPV** was the second best system, as they used a multilingual sentence transformer representation (SBERT) with knowledge distillation. They also introduced an auxiliary language identification task, aside from the downstream check-worthiness task.

**English** Ten teams took part in task 1A for English, submitting a total of 21 runs. The top-ranked team was **NLP&IR@UNED**, and they used several pre-trained transformers models. They reported that **BERTweet was best on the development set**. The model was **trained using RoBERTa on 850 million English tweets and 23 million COVID-19 English tweets**. The second best system (Team **Fight for 4230**) also used BERTweet with a dropout layer; they also used pre-processing and data augmentation.

**Spanish** Six teams took part for Spanish, submitting a total of 13 runs. The top-ranked team **TOBB ETU** explored different data augmentation strategies, including machine translation and weak supervision. Still, their submitted model is a fine-tuned BETO model without data augmentation. The first runner up **GPLSI** opted for the BETO Spanish transformer together with a number of hand-crafted features, such as the occurrence of numbers or words from the LIWC lexicon.

**Turkish** Five teams participated for Turkish, submitting a total of 9 runs. All participants used BERT-based models. The top ranked team **TOBB ETU** fine-tuned BERTurk after removing mentions and URLs. The runner up team **SU-NLP** applied a pre-processing step that includes removing hashtags, emojis, and replacing URLs and mentions with special tokens. Subsequently, they used an ensemble of BERTurk models fine-tuned with different seed values. The third-ranked team **bigIR** machine-translated the Turkish text to Arabic and then fine-tuned AraBERT on the translated text.

Table 8

Subtask 1A: results for the official submissions for all five languages.

	Team	MAP	MRR	RP	P@1	P@3	P@5	P@10	P@20	P@30
Arabic	1 Accenture [71]	0.658	1.000	0.599	1.000	1.000	1.000	1.000	0.950	0.840
	2 bigIR	0.615	0.500	0.579	0.000	0.667	0.600	0.600	0.800	0.740
	3 SCUoL [78]	0.612	1.000	0.599	1.000	1.000	1.000	1.000	0.950	0.780
	4 iCompass	0.597	0.333	0.624	0.000	0.333	0.400	0.400	0.500	0.640
	4 QMUL-SDS [77]	0.597	0.500	0.603	0.000	0.667	0.600	0.700	0.650	0.720
	6 TOBB ETU [80]	0.575	0.333	0.574	0.000	0.333	0.400	0.400	0.500	0.680
	7 DamascusTeam	0.571	0.500	0.558	0.000	0.667	0.600	0.800	0.700	0.640
	8 UPV [81]	0.548	1.000	0.550	1.000	0.667	0.600	0.500	0.400	0.580
	<i>n</i> -gram baseline	0.428	0.500	0.409	0.000	0.667	0.600	0.500	0.450	0.440
Bulgarian	1 bigIR	0.737	1.000	0.632	1.000	1.000	1.000	1.000	1.000	0.800
	2 UPV [81]	0.673	1.000	0.605	1.000	1.000	1.000	1.000	0.800	0.700
	<i>n</i> -gram baseline	0.588	1.000	0.474	1.000	1.000	1.000	0.900	0.750	0.640
	3 Accenture [71]	0.497	1.000	0.474	1.000	1.000	0.800	0.700	0.600	0.440
	4 TOBB ETU [80]	0.149	0.143	0.039	0.000	0.000	0.000	0.200	0.100	0.060
English	1 NLP&IR@UNED [75]	0.224	1.000	0.211	1.000	0.667	0.400	0.300	0.200	0.160
	2 Fight for 4230 [72]	0.195	0.333	0.263	0.000	0.333	0.400	0.400	0.250	0.160
	3 UPV [81]	0.149	1.000	0.105	1.000	0.333	0.200	0.200	0.100	0.120
	4 bigIR	0.136	0.500	0.105	0.000	0.333	0.200	0.100	0.100	0.120
	5 GPLSI [73]	0.132	0.167	0.158	0.000	0.000	0.000	0.200	0.150	0.140
	6 csum112	0.126	0.250	0.158	0.000	0.000	0.200	0.200	0.150	0.160
	7 abaruah	0.121	0.200	0.158	0.000	0.000	0.200	0.200	0.200	0.140
	8 NLytics [84]	0.111	0.071	0.053	0.000	0.000	0.000	0.000	0.050	0.120
	9 Accenture [71]	0.101	0.143	0.158	0.000	0.000	0.000	0.200	0.200	0.100
	10 TOBB ETU [80]	0.081	0.077	0.053	0.000	0.000	0.000	0.000	0.050	0.080
	<i>n</i> -gram baseline	0.052	0.020	0.000	0.000	0.000	0.000	0.000	0.000	0.020
Spanish	1 TOBB ETU [80]	0.537	1.000	0.525	1.000	1.000	0.800	0.900	0.700	0.680
	2 GPLSI [73]	0.529	0.500	0.533	0.000	0.667	0.600	0.800	0.750	0.620
	3 bigIR	0.496	1.000	0.483	1.000	1.000	0.800	0.800	0.600	0.620
	4 NLP&IR@UNED [75]	0.492	1.000	0.475	1.000	1.000	1.000	0.800	0.800	0.620
	5 Accenture [71]	0.491	1.000	0.508	1.000	0.667	0.800	0.900	0.700	0.620
	<i>n</i> -gram baseline	0.450	1.000	0.450	1.000	0.667	0.800	0.700	0.700	0.660
	6 UPV	0.446	0.333	0.475	0.000	0.333	0.600	0.800	0.650	0.580
Turkish	1 TOBB ETU [80]	0.581	1.000	0.585	1.000	1.000	0.800	0.700	0.750	0.660
	2 SU-NLP [79]	0.574	1.000	0.585	1.000	1.000	1.000	0.800	0.650	0.680
	3 bigIR	0.525	1.000	0.503	1.000	1.000	1.000	0.800	0.700	0.720
	4 UPV [81]	0.517	1.000	0.508	1.000	1.000	1.000	1.000	0.850	0.700
	5 Accenture [71]	0.402	0.250	0.415	0.000	0.000	0.400	0.400	0.650	0.660
	<i>n</i> -gram baseline	0.354	1.000	0.311	1.000	0.667	0.600	0.700	0.600	0.460

**All languages.** Table 9 summarizes the MAP performance for all teams that submitted predictions for all languages in Subtask 1A. We can see that team **bigIR** performed best overall.



**Table 9**

**Subtask 1A:** MAP performance for the official submissions for teams participating in all five languages.  $\mu$  shows a standard mean of the five MAP scores;  $\mu_w$  shows a weighed mean, where each MAP is multiplied by the size of the testing set.

	Team	ar	bg	en	es	tr	$\mu$	$\mu_w$
1	bigIR	0.615	<b>0.737</b>	0.136	0.496	0.525	<b>0.502</b>	<b>0.513</b>
2	UPV [81]	0.548	0.673	<b>0.149</b>	0.446	0.517	0.467	0.477
3	TOBB ETU [80]	0.575	0.149	0.081	<b>0.537</b>	<b>0.581</b>	0.385	0.472
4	Accenture [71]	<b>0.658</b>	0.497	0.101	0.491	0.402	0.430	0.456
	<i>n</i> -gram baseline	0.428	0.588	0.052	0.450	0.354	0.374	0.394

#### 4. Subtask 1B: Check-worthiness of debates or speeches.

**Subtask 1B** is a legacy task that has evolved from the first edition of CheckThat! and it was carried over in 2018, 2019, and 2020 [85, 7, 86]. In each edition, more training data from more diverse sources have been added, with all speeches and debates about politics. The task aims to mimic the selection strategy that fact-checking organizations such as *PolitiFact* use to select the sentences and the claims to fact-check. The task is defined as follows:

*“Given a transcript of a speech or a political debate, rank the sentences in the transcript according to the priority with which they should be fact-checked.”*

**Table 10**

**Subtask 1B:** Debate fragments with the check-worthy sentences marked with ✓.

C. Booker:	We have systemic racism that is <u>eroding</u> our nation from health care to the criminal justice system.	✗
C. Booker:	And it’s nice to go all the way back to slavery, but dear God, we have a criminal justice system that is so racially biased, we have more African-Americans under criminal supervision today than all the slaves in 1850.	✓
(a) Fragment from the 2019 Democratic Debate in Detroit		
L. Stahl:	Do you still think that climate change is a hoax?	✓
D. Trump:	I think something’s happening.	✗
D. Trump:	Something’s changing and it’ll change back again.	✗
D. Trump:	I don’t think it’s a hoax, I think there’s probably a difference.	✓
D. Trump:	But I don’t know that it’s manmade.	✓
(b) Fragment from the 2018 CBS’ 60 Minutes interview with President Trump		
D. Trump:	We have no country if we have no border.	✗
D. Trump:	Hillary wants to give <u>amnesty</u> .	✓
D. Trump:	She wants to have open borders.	✓

(c) Fragment from the 2016 third presidential debate

Table 7

Subtask 1B: Statistics about the CT-CWT-21 corpus for subtask 1B.

Dataset	# of debates	# of sentences	
		Check-worthy	Non-check-worthy
Training	40	429	41,604
Development	9	69	3,517
Test	8	298	5,002
Total	57	796	50,123

#### 4.1. Dataset

Often, after a major political event, such as a public debate or a speech by a government official, a professional fact-checker would go through the event transcript and would select a few claims to fact-check. Since those claims were selected for verification, we consider them as check-worthy. This is what we used to collect our data, focusing on *PolitiFact* as a fact-checking source. For a political event (debate/speech), we collected the article from *PolitiFact* and we obtained its official transcript, e.g., from ABC, Washington Post, CSPAN, etc. We then manually matched the sentences from the *PolitiFact* articles to the exact statement that was made in the debate/speech.

We collected a total of 57 debates/speeches from 2012–2018, and we selected sentences from the transcript that were actually fact-checked by human fact-checkers. We relied on *PolitiFact* to identify the sentences from the transcripts that could be fact-checked. As fact-checking is a time-consuming process, *PolitiFact* journalists only fact-check a few claims and there is an abundance of false negative examples in the dataset. Thus, we wanted to address this issue at test time: we manually looked over the debates from the test set, and we attempted to check whether each sentence has a fact-checking verified claim using BM25 suggestions. Table 10 shows some annotated examples, and Table 7 gives some statistics. Note the higher proportion of positive examples in the test set compared to the training and the development sets.

Further details about the construction of the CT-CWT-21 corpus can be found in [87].

#### 4.2. Overview of the Systems

Two teams took part in this subtask, submitting a total of 3 runs. Table 8 shows the performance of the official submissions on the test set, in addition to an  $n$ -gram baseline. Similarly to Task 1A, the official run was the last valid blind submission by each team. The table shows the runs ranked on the basis of the official MAP measure.

The top-ranked team, **Fight for 4230**, fine-tuned BERTweet after normalizing the claims, augmenting the data using WordNet-based substitutions and removal of punctuation. They were able to beat the  $n$ -gram baseline by 18 MAP points absolute. The other team, NLytics [76], fine-tuned RoBERTa, but they could not beat the  $n$ -gram baseline.

**Table 8**

**Task 1B (English):** Official evaluation results, in terms of MAP, MRR, R-Precision, and Precision@ $k$ . The teams are ranked by the official evaluation measure: MAP.

	Team	MAP	MRR	RP	P@1	P@3	P@5	P@10	P@20	P@30
1	Fight for 4230 [72]	0.402	0.917	0.403	0.875	0.833	0.750	0.600	0.475	0.350
	<i>n</i> -gram baseline	0.235	0.792	0.263	0.625	0.583	0.500	0.400	0.331	0.217
2	NLytics [84]	0.135	0.345	0.130	0.250	0.125	0.100	0.137	0.156	0.135

### 4.3. Evaluation

As this task was very similar to Subtask 1A, but for a different genre, we used the same evaluation measures: namely, MAP as the official measure, and we also report P@ $k$  for various values of  $k$ . Table 8 shows the performance of the primary submissions of the participating teams. We can see that the overall results are low, and only one of the teams managed to beat the *n*-gram baseline. Once again, the data and the evaluation scripts are available online.<sup>11</sup>

## 5. Conclusion and Future Work

We have presented an overview of task 1 of the CLEF-2021 CheckThat! lab. The lab featured tasks that span the full verification pipeline: from spotting check-worthy claims to checking whether they have been fact-checked before. Task 1 focused on check-worthiness in tweets about COVID-19 and politics (Subtask 1A), and in political debates and speeches (Subtask 1B). Inline with the general mission of CLEF, we promoted multi-linguality by offering the task in five different languages. The participating systems used transformer models (e.g., BERT and RoBERTa) and some used data augmentation. Applying standard pre-processing was common for many systems, and almost all systems outperformed an *n*-gram baseline for Subtask 1A.

In future work, we plan a new iteration of the CLEF CheckThat! lab and of the task, where we would offer larger training sets, as well as more fine-grained tasks.

## Acknowledgments

The work of Tamer Elsayed, Maram Hasanain and Zien Sheikh Ali was made possible by NPRP grant #NPRP-11S-1204-170060 from the Qatar National Research Fund (a member of Qatar Foundation). The work of Fatima Haouari is supported by GSRA grant #GSRA6-1-0611-19074 from the Qatar National Research Fund. The statements made herein are solely the responsibility of the authors.

This work is also part of the Tanbih mega-project,<sup>12</sup> developed at the Qatar Computing Research Institute, HBKU, which aims to limit the impact of “fake news”, propaganda, and media bias by making users aware of what they are reading, thus promoting media literacy and critical thinking.

<sup>11</sup>[http://gitlab.com/checkthat\\_lab/clef2021-checkthat-lab/](http://gitlab.com/checkthat_lab/clef2021-checkthat-lab/)

<sup>12</sup><http://tanbih.qcri.org>

## References

- [1] P. Nakov, D. S. M. Giovanni, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, B. Hamdan, Z. S. Ali, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, M. Kutlu, Y. S. Kartal, Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, CLEF 2021, 2021.
- [2] S. Shaar, F. Haouari, W. Mansour, M. Hasanain, N. Babulkov, F. Alam, G. Da San Martino, T. Elsayed, P. Nakov, Overview of the CLEF-2021 CheckThat! lab task 2 on detect previously fact-checked claims in tweets and political debates, in: [88], 2021.
- [3] G. K. Shahi, J. M. Struß, T. Mandl, Overview of the CLEF-2021 CheckThat! lab: Task 3 on fake news detection, in: [88], 2021.
- [4] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, S. Shaar, Z. Sheikh Ali, Overview of CheckThat! 2020: Automatic identification and verification of claims in social media, LNCS (12260), 2020.
- [5] T. Elsayed, P. Nakov, A. Barrón-Cedeño, M. Hasanain, R. Suwaileh, G. Da San Martino, P. Atanasova, Overview of the CLEF-2019 CheckThat!: Automatic identification and verification of claims, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, LNCS, 2019, pp. 301–321.
- [6] P. Atanasova, L. Màrquez, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, W. Zaghoulani, 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: *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, 2018.
- [7] P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, G. Da San Martino, Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 1: Check-worthiness, in: [89], 2019.
- [8] A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, P. Atanasova, W. Zaghoulani, S. Kyuchukov, G. Da San Martino, P. Nakov, Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims, task 2: Factuality, in: *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, 2018.
- [9] T. Elsayed, P. Nakov, A. Barrón-Cedeño, M. Hasanain, R. Suwaileh, G. Da San Martino, P. Atanasova, CheckThat! at CLEF 2019: Automatic identification and verification of claims, in: L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, D. Hiemstra (Eds.), *Advances in Information Retrieval, ECIR '19*, 2019, pp. 309–315.
- [10] M. Hasanain, R. Suwaileh, T. Elsayed, A. Barrón-Cedeño, P. Nakov, Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 2: Evidence and factuality, in: [89], 2019.
- [11] P. Nakov, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghoulani, P. Atanasova, S. Kyuchukov, G. Da San Martino, Overview of the CLEF-2018 Check-That! lab on automatic identification and verification of political claims, in: *Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets*

Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science, 2018, pp. 372–387.

- [12] P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, I. Koychev, A context-aware approach for detecting worth-checking claims in political debates, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, 2017, pp. 267–276.
- [13] N. Hassan, C. Li, M. Tremayne, Detecting check-worthy factual claims in presidential debates, in: J. Bailey, A. Moffat, C. C. Aggarwal, M. de Rijke, R. Kumar, V. Murdock, T. K. Sellis, J. X. Yu (Eds.), Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM, 2015, pp. 1835–1838.
- [14] I. Jaradat, P. Gencheva, A. Barrón-Cedeño, L. Màrquez, P. Nakov, ClaimRank: Detecting check-worthy claims in Arabic and English, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2018, pp. 26–30.
- [15] S. Vasileva, P. Atanasova, L. Màrquez, A. Barrón-Cedeño, P. Nakov, It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP '19, 2019, pp. 1229–1239.
- [16] S. Shaar, N. Babulkov, G. Da San Martino, P. Nakov, That is a known lie: Detecting previously fact-checked claims, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3607–3618.
- [17] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, G. D. S. Martino, Automated fact-checking for assisting human fact-checkers (2021).
- [18] S. Shaar, F. Alam, G. D. S. Martino, P. Nakov, The role of context in detecting previously fact-checked claims, arXiv:2104.07423 (2021).
- [19] I. Augenstein, C. Lioma, D. Wang, L. Chaves Lima, C. Hansen, C. Hansen, J. G. Simonsen, MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, 2019, pp. 4685–4697.
- [20] G. Karadzhov, P. Nakov, L. Màrquez, A. Barrón-Cedeño, I. Koychev, Fully automated fact checking using external sources, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, 2017, pp. 344–353.
- [21] M. Mohtarami, R. Baly, J. Glass, P. Nakov, L. Màrquez, A. Moschitti, Automatic stance detection using end-to-end memory networks, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, 2018, pp. 767–776.
- [22] M. Mohtarami, J. Glass, P. Nakov, Contrastive language adaptation for cross-lingual stance detection, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, 2019, pp. 4442–4452.
- [23] P. Atanasova, P. Nakov, L. Màrquez, A. Barrón-Cedeño, G. Karadzhov, T. Mihaylova, M. Mohtarami, J. Glass, Automatic fact-checking using context and discourse information,

Journal of Data and Information Quality (JDIQ) 11 (2019) 12.

- [24] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, R. Kumar (Eds.), Proceedings of the 20th International Conference on World Wide Web, WWW 2011, 2011, pp. 675–684.
- [25] D. Kopev, A. Ali, I. Koychev, P. Nakov, Detecting deception in political debates using acoustic and textual features, in: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, ASRU '19, 2019, pp. 652–659.
- [26] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2931–2937.
- [27] R. Baly, G. Karadzhov, D. Alexandrov, J. Glass, P. Nakov, Predicting factuality of reporting and bias of news media sources, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3528–3539.
- [28] R. Baly, M. Mohtarami, J. Glass, L. Màrquez, A. Moschitti, P. Nakov, Integrating stance detection and fact checking in a unified corpus, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 21–27.
- [29] V. Nguyen, K. Sugiyama, P. Nakov, M. Kan, FANG: leveraging social context for fake news detection using graph representation, in: M. d'Aquin, S. Dietze, C. Hauff, E. Curry, P. Cudré-Mauroux (Eds.), CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19–23, 2020, 2020, pp. 1165–1174.
- [30] K. Popat, S. Mukherjee, J. Strötgen, G. Weikum, Where the truth lies: Explaining the credibility of emerging claims on the web and social media, in: Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17, 2017, pp. 1003–1012.
- [31] 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), 2018, pp. 809–819.
- [32] A. Patwari, D. Goldwasser, S. Bagchi, TATHYA: A multi-classifier system for detecting check-worthy statements in political debates, in: E. Lim, M. Winslett, M. Sanderson, A. W. Fu, J. Sun, J. S. Culpepper, E. Lo, J. C. Ho, D. Donato, R. Agrawal, Y. Zheng, C. Castillo, A. Sun, V. S. Tseng, C. Li (Eds.), Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM, 2017, pp. 2259–2262.
- [33] R. Agez, C. Bosc, C. Lespagnol, J. Mothe, N. Petitcol, IRIT at CheckThat! 2018, in: [90], 2018.
- [34] B. Ghanem, M. Montes-y Gómez, F. Rangel, P. Rosso, UPV-INAOE-Autoritas - Check That: Preliminary approach for checking worthiness of claims, in: [90], 2018.
- [35] C. Hansen, C. Hansen, J. Simonsen, C. Lioma, The Copenhagen team participation in the check-worthiness task of the competition of automatic identification and verification of claims in political debates of the CLEF-2018 fact checking lab, in: [90], 2018.
- [36] C. Zuo, A. Karakas, R. Banerjee, A hybrid recognition system for check-worthy claims using heuristics and supervised learning, in: [90], 2018.
- [37] B. Altun, M. Kutlu, TOBB-ETU at CLEF 2019: Prioritizing claims based on check-worthiness, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and



Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.

- [38] L. Coca, C.-G. Cusmuluc, A. Iftene, CheckThat! 2019 UAICS, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.
- [39] R. Dhar, S. Dutta, D. Das, A hybrid model to rank sentences for check-worthiness, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.
- [40] L. Favano, M. Carman, P. Lanzi, TheEarthIsFlat's submission to CLEF'19 CheckThat! challenge, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.
- [41] J. Gasior, P. Przybyła, The IPIAN team participation in the check-worthiness task of the CLEF2019 CheckThat! lab, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.
- [42] C. Hansen, C. Hansen, J. Simonsen, C. Lioma, Neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.
- [43] S. Mohtaj, T. Himmelsbach, V. Woloszyn, S. Möller, The TU-Berlin team participation in the check-worthiness task of the CLEF-2019 CheckThat! lab, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.
- [44] T. Su, C. Macdonald, I. Ounis, Entity detection for check-worthiness prediction: Glasgow Terrier at CLEF CheckThat! 2019, in: CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.
- [45] J. Martinez-Rico, L. Araujo, J. Martinez-Romo, NLP&IR@UNED at CheckThat! 2020: A preliminary approach for check-worthiness and claim retrieval tasks using neural networks and graphs, in: [91], 2020.
- [46] T. McDonald, Z. Dong, Y. Zhang, R. Hampson, J. Young, Q. Cao, J. Leidner, M. Stevenson, The University of Sheffield at CheckThat! 2020: Claim identification and verification on Twitter, in: [91], 2020.
- [47] Y. S. Kartal, M. Kutlu, TOBB ETU at CheckThat! 2020: Prioritizing English and Arabic claims based on check-worthiness, in: [91], 2020.
- [48] F. Alam, S. Shaar, F. Dalvi, H. Sajjad, A. Nikolov, H. Mubarak, G. Da San Martino, A. Abdelali, N. Durrani, K. Darwish, P. Nakov, Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society, ArXiv preprint 2005.00033 (2020).
- [49] F. Alam, F. Dalvi, S. Shaar, N. Durrani, H. Mubarak, A. Nikolov, G. Da San Martino, A. Abdelali, H. Sajjad, K. Darwish, P. Nakov, Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 15, 2021, pp. 913–922.
- [50] E. Williams, P. Rodrigues, V. Novak, Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models, in: [91], 2020.
- [51] M. Hasanain, T. Elsayed, bigIR at CheckThat! 2020: Multilingual BERT for ranking Arabic tweets by check-worthiness, in: [91], 2020.

- [52] G. S. Cheema, S. Hakimov, R. Ewerth, Check\_square at CheckThat! 2020: Claim detection in social media via fusion of transformer and syntactic features, in: [91], 2020.
- [53] A. Hussein, A. Hussein, N. Ghneim, A. Joukhadar, DamascusTeam at CheckThat! 2020: Check worthiness on Twitter with hybrid CNN and RNN models, in: [91], 2020.
- [54] I. Touahri, A. Mazroui, EvolutionTeam at CheckThat! 2020: Integration of linguistic and sentimental features in a fake news detection approach, in: [91], 2020.
- [55] R. Alkhalifa, T. Yoong, E. Kochkina, A. Zubiaga, M. Liakata, QMUL-SDS at CheckThat! 2020: Determining COVID-19 tweet check-worthiness using an enhanced CT-BERT with numeric expressions, in: [91], 2020.
- [56] A. Nikolov, G. Da San Martino, I. Koychev, P. Nakov, Team\_Alex at CheckThat! 2020: Identifying check-worthy tweets with transformer models, in: [91], 2020.
- [57] C.-G. Cusmuluic, L.-G. Coca, A. Iftene, UAICS at CheckThat! 2020: Fact-checking claim prioritization, in: [91], 2020.
- [58] S. Krishan T, K. S, T. D, R. Vardhan K, A. Chandrabose, Tweet check worthiness using transformers, CNN and SVM, in: [91], 2020.
- [59] A. Gupta, P. Kumaraguru, C. Castillo, P. Meier, TweetCred: Real-time credibility assessment of content on Twitter, in: Proceeding of the 6th International Social Informatics Conference, SocInfo '14, 2014, pp. 228–243.
- [60] T. Mitra, E. Gilbert, CREDBANK: A large-scale social media corpus with associated credibility annotations, in: Proceedings of the Ninth International AAAI Conference on Web and Social Media, ICWSM '15, 2015, pp. 258–267.
- [61] A. Gupta, V. Sri Kumar, X-fact: A new benchmark dataset for multilingual fact checking, ArXiv:2106.09248 (2021).
- [62] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, SIGKDD Explor. Newsl. 19 (2017) 22–36.
- [63] Z. Zhao, P. Resnick, Q. Mei, Enquiring minds: Early detection of rumors in social media from enquiry posts, in: A. Gangemi, S. Leonardi, A. Panconesi (Eds.), Proceedings of the 24th International Conference on World Wide Web WWW, 2015, pp. 1395–1405.
- [64] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, A. Scala, The COVID-19 social media infodemic, arXiv:2003.05004 (2020).
- [65] X. Song, J. Petrak, Y. Jiang, I. Singh, D. Maynard, K. Bontcheva, Classification aware neural topic model and its application on a new COVID-19 disinformation corpus, arXiv:2006.03354 (2020).
- [66] X. Zhou, A. Mulay, E. Ferrara, R. Zafarani, Recovery: A multimodal repository for COVID-19 news credibility research, in: M. d'Aquin, S. Dietze, C. Hauff, E. Curry, P. Cudré-Mauroux (Eds.), CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19–23, 2020, 2020, pp. 3205–3212.
- [67] F. Haouari, M. Hasanain, R. Suwaileh, T. Elsayed, ArCOV-19: The first Arabic COVID-19 Twitter dataset with propagation networks, arXiv preprint arXiv:2004.05861 (2020).
- [68] F. Haouari, M. Hasanain, R. Suwaileh, T. Elsayed, ArCOV19-rumors: Arabic COVID-19 Twitter dataset for misinformation detection, arXiv preprint arXiv:2010.08768 (2020).
- [69] L. Konstantinovskiy, O. Price, M. Babakar, A. Zubiaga, Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection, arXiv:1809.08193 (2018).

- [70] Y. S. Kartal, M. Kutlu, TrClaim-19: The first collection for Turkish check-worthy claim detection with annotator rationales, in: Proceedings of the 24th Conference on Computational Natural Language Learning, 2020, pp. 386–395.
- [71] E. Williams, P. Rodrigues, S. Tran, Accenture at CheckThat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation, in: [88], 2021.
- [72] X. Zhou, B. Wu, P. Fung, Fight for 4230 at CLEF CheckThat! 2021: Domain-specific preprocessing and pretrained model for ranking claims by check-worthiness, in: [88], 2021.
- [73] R. Sepúlveda-Torres, E. Saquete, GPLSI team at CLEF CheckThat! 2021: Fine-tuning BETO and RoBERTa, in: [88], 2021.
- [74] O. Rjab, H. Haddad, W. Henia, C. Fourati, iCompass at CheckThat! 2021: Identifying check-worthy Arabic tweets, in: [88], 2021.
- [75] J. M.-R. Juan R. Martinez-Rico, L. Araujo, NLP&IR@UNED at CheckThat! 2021: Check-worthiness estimation and fake news detection using transformer models, in: [88], 2021.
- [76] A. Pritzkau, NLytics at CheckThat! 2021: Check-worthiness estimation as a regression problem on transformers, in: [88], 2021.
- [77] A. S. Abumansour, A. Zubiaga, QMUL-SDS at CheckThat! 2021: Enriching pre-trained language models for the estimation of check-worthiness of Arabic tweets, in: [88], 2021.
- [78] S. Alhabiti, M. Alsalka, E. Atwell, An AraBERT model for check-worthiness of Arabic tweets, in: [88], 2021.
- [79] B. Carik, R. Yeniterzi, SU-NLP at CheckThat! 2021: Check-worthiness of Turkish tweets, in: [88], 2021.
- [80] M. S. Zengin, Y. S. Kartal, M. Kutlu, TOBB ETU at CheckThat! 2021: Data engineering for detecting check-worthy claims, in: [88], 2021.
- [81] I. Baris Schlicht, A. Magnossão de Paula, P. Rosso, UPV at CheckThat! 2021: Mitigating cultural differences for identifying multilingual check-worthy claims, in: [88], 2021.
- [82] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, 2019, pp. 4171–4186.
- [83] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, Arxiv:1907.11692 (2019).
- [84] A. Pritzkau, NLytics at CheckThat! 2021: Multi-class fake news detection of news articles and domain identification with RoBERTa - a baseline model, in: [88], 2021.
- [85] P. Atanasova, L. Marquez, 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: [90], 2018.
- [86] 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: [91], 2020.
- [87] 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: [88], 2021.

- [88] G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Working Notes. Working Notes of CLEF 2021–Conference and Labs of the Evaluation Forum, 2021.
- [89] L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2019.
- [90] L. Cappellato, N. Ferro, J.-Y. Nie, L. Soulier (Eds.), Working Notes of CLEF 2018–Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2018.
- [91] L. Cappellato, C. Eickhoff, N. Ferro, A. Névél (Eds.), CLEF 2020 Working Notes, CEUR Workshop Proceedings, 2020.