

Overview of the CLEF-2021 CheckThat! Lab: Task 3 on Fake News Detection

Gautam Kishore Shahi¹, Julia Maria Struß² and Thomas Mandl³

¹University of Duisburg-Essen, Germany

²Information Science, Potsdam University of Applied Sciences, Germany

³Information Science, University of Hildesheim, Germany

Abstract

We describe the fourth edition of the CheckThat! Lab, part of the 2021 Conference and Labs of the Evaluation Forum (CLEF). The lab evaluates technology supporting three tasks related to factuality, and it covers Arabic, Bulgarian, English, Spanish, and Turkish. Here, we present *task 3*, which focuses on multi-class fake news detection and topical domain detection of news articles. Overall, there were 88 submissions by 27 teams for Task 3A, and 49 submissions by 20 teams for task 3B (two team from Task 3A and seven teams from Task 3B are excluding from the ranking due to wrong submission file). The best performing system for task 3A achieved a macro F_1 -score of 0.84 and was ahead of the rest by a rather large margin. The performance of the systems for task 3B was overall higher than for task 3A with the top performing system achieving a macro F_1 -score of 0.88. In this paper, we describe the process of data collection and the task setup, including the evaluation measures used, and we give a brief overview of the participating systems. Last but not least, we release to the research community all data sets from the lab as well as the evaluation scripts, which should enable further research in automatic classification of news articles with respect to their correctness and topical domain.

Keywords

Misinformation, Fake news, Text classification, Evaluation, Deep learning, Domain Identification

1. Introduction

Misinformation is a huge societal problem, and it appears in many different forms [1]. Often, a part of a true story is left out, or something is added to create misleading articles. Misinformation opens much demand for future research [2]. Technology for identification and evaluation is necessary to support the effort to eliminate wrong and misleading information. In the long term, misinformation may damage the trust in media and create harm for the discourse within society.

CheckThat! at CLEF is a lab that intends to evaluate system to perform fake news identification and provide tools for checking pipeline supporting humans. This lab consists of three tasks which are all described in the overview paper [3, 4]. Much work was recently dedicated towards the identification of misinformation in general [5, 6, 7] and in particular in social media [8].

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

✉ gautam.shahi@uni-due.de (G. K. Shahi); struss@fh-potsdam.de (J. M. Struß); mandl@uni-hildesheim.de (T. Mandl)

ORCID 0000-0001-6168-0132 (G. K. Shahi); 0000-0001-9133-4978 (J. M. Struß); 0000-0002-8398-9699 (T. Mandl)



© 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)

The dissemination of misinformation in social media gives users more hints on the lack of quality because the source of information remains unclear. However, misinformation spread in venues that resemble genuine news outlets or appear even in usually reliable newspapers poses a big threat. Users trust such outlets and sources and might not be suspicious. Much of the research on automatic identification of fake news has been dedicated to the analysis of social media. Data collections often are built from social media platforms or sources like Wikipedia. Data collections that assemble news articles and classify them into misinformation and genuine information rare. In task 3 of CheckThat! 2021, we approached the problem of creating such a resource and providing it for a shared task. The shared task is organised based on the news articles, and the goal was to predict the truthfulness of articles and their categorical domain. A detailed description of the task is given in section 3.

The remainder of the paper is organised as follows: section 2 describes the state of the art research, section 3 provides the task descriptions, section 4 emphasis on the steps involved in the data collection, section 5 focuses on the submissions and results obtained from the participants, section 6 provides a detailed description of the different approaches used the by participants. In the end, we provide a brief conclusion and an outlook on potential future work in section 7.

2. Related Work

Determining the credibility of a claim is a research problem, that attracted significant attention in recent years [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 12, 22]. Even though check worthy claims can come from a vast variety of different sources, most work focuses on those from social media [23, 24, 25, 26]. The verification of claims in news articles has also attracted some attention but has been addressed as a binary classification problem mostly [27, 28]. Still, for partially false news, different patterns of circulation have been observed, and false news propagates faster than partially false news on social media [29, 30].

Besides the CheckThat! lab at CLEF, there have been several initiatives at different evaluation campaigns like SemEval in the last years approaching different aspects or sources of claims: RumourEval focused on determining the truthfulness of rumours [31, 32], other tasks at SemEval addressed stance [33] and propaganda detection [34], or fact-checking in community question answering forums [35].

Other initiatives include the FakeNews task at MediaEval [36] dealing especially with false information concerning coronavirus and 5G conspiracy theories, or the FEVER task [37] on fact extraction and verification employing unstructured and structured evidence from Wikipedia.

So far, mainly Wikipedia [38] and social media posts have been used to create collections for the benchmark (e.g. [39]) studies.

Disinformation is also often transported via images, videos or other non-textual material. A dataset addressing this facet of the problem has been proposed by [40] with the *Fakeddit* corpus providing not only text but also images from Reddit. In another study, the author have analysed the spread of misinformation on the WhatsApp tiplines in the context of fact-checking [41]

3. Task Description

Task 3 is divided in two subtasks both of which are classification tasks and were running for the first time as a pilot in this years' CheckThat! iteration. They were offered in English.

Subtask 3A: Multi-class fake news detection of news articles. Given the text and title of a news article, determine whether the main claim made in the article is *true*, *partially true*, *false*, or *other*. The four categories were proposed based on Shahi et al. [29, 42] and the definitions for the four categories being as follows:

False The main claim made in an article is untrue.

Partially False The main claim of an article is a mixture of true and false information. It includes articles in categories like partially false, partially true, mostly true, miscaptioned, misleading etc., as defined by different fact-checking services.

True This rating indicates that the primary elements of the main claim are demonstrably true.

Other An article that cannot be categorised as true, false, or partially false due to lack of evidence about its claims. This category includes articles in dispute and unproven articles.

Subtask 3B: Topical domain detection of news articles. Given the text of a news article, determine the topical domain (health, crime, climate, election, and education) of the article.

4. Data Description

Fact-checking of claims made in news articles or social media is a time-consuming process, requiring substantial background knowledge on the claims' topic as well as sufficient experience and training on the task. To be able to provide data from a broad range of topics, we relied on the judgements of experts in the field, provided through relevant fact-checking services on their respective websites. An overview of the applied approach is given in Figure 1 and described in more detail in the following sub-sections.

4.1. Crawling Fact-Checking Reports

In a first step, fact-checking websites were identified and analysed with regard to their respective structures. The published fact-checking reports differ substantially regarding format and content. Therefore, tailored crawlers and scrapers for each of the considered sites were necessary. The entire crawling process was based on the AMUSED framework [43]. An overview of the respective websites and the data collected from each individual site can be found in Table 1.

From each fact-checking report, we collected the claims together with the experts' judgement on the correctness of the respective claims as well as the links to the potential original sources of the claims, and if available, information on the type of the source (e. g. news article, social media posts, images) as well as the topical domain of the article. Some of the websites offered

Creating a Dataset of Original misinformation articles

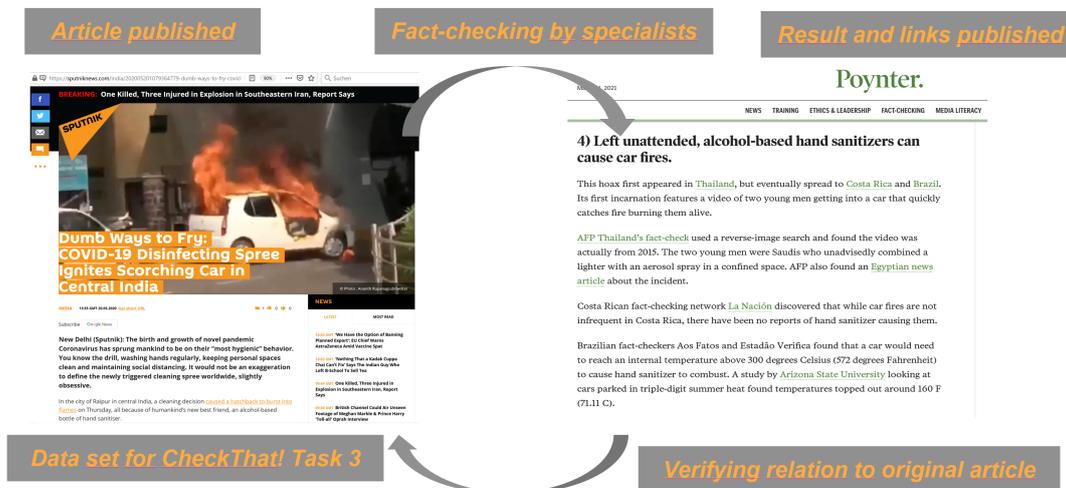


Figure 1: Overview of data crawling from fact-checked articles.

these details as metadata in JSON format using the *ClaimReview*-type defined by *Schema.org*. Even though the metadata was not always complete for claims coming from those reports, the link to the original source could be identified reliably. However, for fact-checking reports not providing corresponding metadata a manual check of the different links given in the reports was necessary as most links in fact-checking reports point to sources supporting the judgement on the claim's correctness and generally no clear position of the link to the claim's source could be identified. To keep the task manageable, only the first three links provided in each report was considered due to the observation that most of the time, one of these referred to the claim's source page.

Before conducting the manual source identification step, automatic filtering was applied, removing all claims stemming from social media posts or multimedia documents. These could be identified due to given source type information or the domains' given in the links. This left about 3,400 article candidates for manual checking of the more than 200,000 crawled fact-checking reports.

Besides the correct identification of the source for each entry, we also manually verified that the source was indeed an article and that the article was still available on the site, thus eliminating error pages or articles with different content than at the time of fact-checking. More than half of the article candidates were removed due to this filtering, resulting in a set of about 1,400 articles for the task.

Table 1

Details on the number article candidates and articles obtained from the reports available on the websites of different fact-checking services

fact-checking service	#article candidates	#articles
aap	15	4
afp	53	33
altnews	5	3
checkyourfact	4	0
climatefeedback	248	174
dubawa	32	21
factcheckni	15	8
fullfact	440	293
healthfeedback	180	139
leadstories	319	236
mythdetector	23	17
politifact	1,660	337
polygraph	175	75
theferret	17	10
truthorfiction	192	77
sum	3378	1427

4.2. Scraping Articles

For each of the remaining articles, title and text were extracted from the respective pages in an automatic scraping process. The multitude of different news websites did not allow for tailored scrapers. Hence we extracted h1-tags as titles and the content of p-tags as text, excluding content from footers.

To ensure high data quality, articles with missing titles or text were again checked manually. The same is true for articles with a text length below a threshold of 500 characters. These articles often were not extracted correctly, sitting behind a paywall, or removed in the short period between the annotation and extraction process. Whenever possible, we added the missing content manually; otherwise, the article was removed from the corpus.

4.3. Data Set for Task 3A

As described above, we relied on the judgements of fact-checking experts for task 3a. However, due to the heterogeneous labelling schemes of different fact-checking agencies (e.g., false: incorrect, inaccurate, misinformation), we did not provide the original labels. Instead, labels with a joint meaning were merged and sometimes grouped under a concept with a broader meaning, as described by [29, 44]. Some examples are given in the following table.

The participants were provided with a training data set consisting of 900 news articles, leaving 354 articles for testing. The statistics about the class distributions in the training and test data can be found in Table 4. Besides a unique identifier for each article, only the title and text, as well as the class label, were provided. No further metadata was included. Table 3 shows some sample data.

Table 2

Examples for labels merged to build the final set of classes for task 3a

task label	original label
false	fake, false, inaccurate, inaccurate with consideration, incorrect, not true, pants-fire
partially false	half-true, imprecise, mixed, partially true, partly true, barely-true, misleading
true	accurate, correct, true
other	debated, other, unclear, unknown, unsupported

Table 3

Sample data for task 3a

public_id	title	text	our rating
c5175d8d	Paul Ryan’s Worst Ally - The New York Times	WHATEVER drama plays out when Republicans meet in Cleveland next week to nominate their party’s presidential candidate, the most consequential story line might well be the nationally televised debut of the awkward political partnership [...]	true
392886ea	Antifa gearing up for false flag violence disguised as Trump-supporters	With merchants in Democrat-run cities boarding up their storefront windows, the possibility of serious urban violence is well understood. As horrific as that could get (ask anyone who lives in Minneapolis), the progressive fascists appear to have a plan in place to make it even worse: [...]	false

4.4. Data Set for Task 3B

A subset of 455 articles was annotated with their respective topic, based on information provided by the fact-checking services, a challenge being the variety of topics in the data set and thus small numbers of examples. Individual topics were grouped to broader topical domains including *health*, *climate*, *economy*, *crime*, *elections*, and *education*. Altogether, 318 articles were provided as training data, leaving 137 articles for testing. The respective class distributions are shown in Table 4.

The complete CT-FAN-21 corpus used for tasks 3A and 3B is available at Zenodo [45].

5. Submissions and Results

In this section, we present an overview of all task submissions for tasks 3A and 3B. Overall, there were 88 submissions by 27 teams for Task 3A and 49 submissions by 20 teams for task 3B. Each participating group could submit up to five runs in both of the subtasks. After evaluation, we found that two teams from task 3A and seven teams from task 3B submitted files not matching

Table 4

Task 3: Statistics about the number of documents and class distribution for the CT-FAN-21 corpus for fake news detection (left) and for topic identification (right).

Class	Training	Test	Topic	Training	Test
False	465	111	Health	127	54
True	142	65	Climate	49	21
Partially false	217	138	Economy	43	19
Other	76	40	Crime	39	17
Total	900	354	Elections	32	14
			Education	28	12
			Total	318	137

the specified format, thus having been eliminated from the evaluation. In Tables 5 and 6 the best submission of each team for task 3A and 3B are given, respectively.

Both tasks 3A and 3B are classification tasks; therefore we used accuracy and macro- F_1 score for evaluation, ranking systems by the latter.

5.1. Multi-class fake news categorization of news articles

Most teams used deep learning models, and in particular, transformer architectures were applied successfully for the task. This task being a pilot, there have been no attempts to model knowledge with semantic technology, e.g., argument processing [61].

The best run for this task was submitted by team **NoFake**. They were ahead of the rest by a substantial margin and achieved a macro- F_1 score of 0.838. The team applied BERT and made extensive use of external resources. In particular, they downloaded collections of misinformation datasets from fact-checking sites as additional training data. The second best submission (team **Saud**) achieved a macro- F_1 score of 0.503. The team used lexical features, traditional weighting methods as features, and standard machine learning algorithms. Traditional approaches can still outperform deep learning models for this task.

Many teams used BERT and its newer variants, often without fine-tuning. The most popular model was RoBERTa, which was used by seven teams. Team **MUCIC** used a majority voting ensemble with three BERT variants [58]. The participating teams that used BERT had to tackle the issue of handling the length of the input: BERT and its variants have limitations for the length of the input. However, the length of texts in the CT-FAN-21 dataset, which consists of newspaper articles, is often longer. In most cases, heuristics were used for the selection of a part of the article text. Overall, most submissions achieved a macro- F_1 score below 0.5.

A more detailed discussion of the different approaches is given in section 6.

5.2. Topical domain identification of news articles

The performance of the systems for task 3B was overall higher than for task 3A. The three best submissions achieved a similar performance, and they all applied transformer-based architectures. The best submission, by team **NITK_NLP**, used an ensemble of three transformers [48].

Table 5

Performance of the best run per team for **task 3A** based on F_1 score for individual classes, and accuracy and macro- F_1 for the overall measure.

Team	True	False	Partially False	Other	Accuracy	Macro-F1
1 NoFake* [46]	0.824	0.862	0.879	0.785	0.853	0.838
2 Saud*	0.321	0.615	0.502	0.618	0.537	0.514
3 DLRG*	0.250	0.588	0.519	0.656	0.528	0.503
4 NLP&IR@UNED [47]	0.247	0.629	0.536	0.459	0.528	0.468
5 NITK_NLP [48]	0.196	0.617	0.523	0.459	0.517	0.449
6 UAICS [49]	0.442	0.470	0.482	0.391	0.458	0.446
7 CIVIC-UPM [50]	0.268	0.577	0.472	0.340	0.463	0.414
8 Uni. Regensburg [51]	0.231	0.489	0.497	0.400	0.438	0.404
9 Pathfinder* [52]	0.277	0.517	0.451	0.360	0.452	0.401
10 CIC* [53]	0.205	0.542	0.490	0.319	0.410	0.389
11 Black Ops [54]	0.231	0.518	0.327	0.453	0.427	0.382
12 NLytics*	0.130	0.575	0.522	0.318	0.475	0.386
13 Nkovachevich [55]	0.237	0.643	0.552	0.000	0.489	0.358
14 talhaanwar*	0.283	0.407	0.435	0.301	0.367	0.357
15 abaruah	0.165	0.531	0.552	0.125	0.455	0.343
16 Team GPLSI[56]	0.293	0.602	0.226	0.092	0.356	0.303
17 Sigmoid [57]	0.222	0.345	0.323	0.154	0.291	0.261
18 architap	0.154	0.291	0.394	0.187	0.294	0.257
19 MUCIC [58]	0.143	0.446	0.275	0.070	0.331	0.233
20 Probity	0.163	0.401	0.335	0.033	0.302	0.233
21 M82B [59]	0.130	0.425	0.241	0.094	0.305	0.223
22 Spider	0.046	0.482	0.145	0.069	0.316	0.186
23 Qword [60]	0.108	0.458	0.000	0.033	0.277	0.150
24 ep*	0.060	0.479	0.000	0.000	0.319	0.135
25 azaharudue*	0.060	0.479	0.000	0.000	0.319	0.135
<i>Majority class baseline</i>	0.000	0.477	0.000	0.000	0.314	0.119

* Runs submitted after the deadline, but before the release of the results.

The second best submission (by team **NoFake**) and the third best submission (by team **Nkovachevich**) used BERT.

6. Discussion of the Approaches Used

The participants have used a large variety of resources and models. Machine learning and pre-trained deep models were applied most frequently. The good performance of deep learning models is in line with the results of other recent shared tasks in text classification on complex tasks like hate speech identification [62], sexism detection [63] or sentiment analysis [64].

Table 6

Performance of the best run per team for **task 3B** based on F1-measure for individual classes, and accuracy and macro-F₁ for overall measure.

Team	Climate	Crime	Economy	Education	Elections	Health	Acc	Macro F1
1 NITK_NLP [48]	0.950	0.872	0.824	0.800	0.897	0.946	0.905	0.881
2 NoFake* [46]	0.800	0.875	0.900	0.957	0.692	0.907	0.869	0.855
3 Nkovachevich [55]	0.927	0.872	0.743	0.737	0.857	0.911	0.869	0.841
4 DLRG	0.952	0.743	0.688	0.800	0.828	0.897	0.847	0.818
5 CIC* [53]	0.952	0.750	0.688	0.588	0.889	0.871	0.832	0.790
6 architap	0.900	0.711	0.774	0.609	0.815	0.907	0.825	0.786
7 NLytics	0.826	0.714	0.710	0.500	0.769	0.867	0.788	0.731
8 CIVIC-UPM* [50]	0.864	0.700	0.645	0.421	0.609	0.821	0.745	0.677
9 ep*	0.727	0.476	0.222	0.343	0.545	0.561	0.511	0.479
10 Pathfinder* [52]	0.900	0.348	0.250	0.000	0.526	0.667	0.599	0.448
11 M82B [59]	0.294	0.000	0.000	0.000	0.000	0.576	0.409	0.145
12 MUCIC [58]	0.294	0.000	0.000	0.000	0.000	0.576	0.409	0.145
13 azaharudue*	0.129	0.000	0.000	0.125	0.000	0.516	0.321	0.128
<i>Majority class baseline</i>	0.000	0.000	0.000	0.000	0.000	0.565	0.394	0.094

* Runs submitted after the deadline, but before the release of the results.

6.1. Classification Approaches

Many teams have used deep learning models, and they have chosen in particular transformer architectures that have been successful on many classification tasks recently. There have been no attempts to model knowledge with so-called semantic technology (e.g. transforming text into triplets [65]) or to extract statements that could be checked against knowledge bases (e.g. argument retrieval [61]). Such technologies could also be potentially useful as technology for detecting misinformation.

Most teams applied neural network based transformers. They often relied on BERT and its newer variants. The most popular model used was RoBERTa: [55, 49, 58, 66, 50, 48].

One team used a majority voting ensemble with 3 BERT variants [58].

The typical approaches included the download of a pre-trained model, and it is further fine-tuning using the data provided. The most popular pre-trained model was Glove [67].

Fine-tuning details were not always fully provided when BERT or variants were used.

The participating teams using BERT had to find solutions for the length of the input. Typically, BERT and its variants are used for shorter text. However, the length of articles in the newspaper dataset provided is much longer. Heuristics were mainly used for the selection of a part of the text. Several approaches used at the beginning of the article.

Also, other neural network models were used; in particular, recurrent neural networks were also popular. Team Kovachevich applied a LSTM based on GLOVE embeddings [55]. LSTMs were used by team Spider and several other teams [57, 68] and Bi-LSTM only by one team [59]. One approach combined a LSTM and a Bi-LSTM [49].

Often teams experimented also with traditional text processing methods as they are used for knowledge representation in information retrieval. TFIDF models were used. For example, Kovachevich used a Naïve Bayes classifier for TF-IDF features for the 500 most frequent stems in the dataset [55].

Lexical features in combination with bi-grams and POS as more complex language features were also used [49].

6.2. Additional Approaches

Some teams used additional processing techniques which are not part of standard text classification algorithms. The Linguistic inquiry and word count (LIWC) tool has been applied [47]. LIWC intends to extract features characterising the personality of a writer. Another lexical approach is taken by the Stanford Empath Tool, which can be used for semantic analysis. It was also applied to create further features for the classification [49]. The massive use of external resources led to good success for the team NoFake. An interesting data augmentation technique was suggested. It included inserting artificially created documents that were similar to the training documents [53]. However, it did not lead to very good results.

The idea of fake news being shared by less authoritative sites has been applied as well. One group implemented an authority model or proxy. They used the title as a Web search engine query and checked the authority of the sites which appeared in the hit list. The appearance of fact-checking sites was considered as a negative indicator for the misinformation status [47]. Some of the authors mentioned including the temporal information for classification as described in [69, 70]. Also, it would be beneficial for an agent if we include the information about the fact-checked articles in HTML markup [71].

6.3. Detailed Description of Participants Systems

In this subsection, we provide a detailed description of the individual participant papers to offer deeper insight into the individual approaches applied to the tasks.

Team Black Ops [54] (3A:11) performed data pre-processing by removing stop-words and punctuation marks. Then, they experimented with decision trees, random forest, and gradient boosting classifiers for Task 3A and found the latter to perform best.

Team CIC [53] (3A:10 3B:5) experimented with logistic regression, multi-layer perceptron, support vector machines, and random forest. Their experiments consisted of using stratified 5-fold cross-validation on the training data. Their best results were obtained using logistic regression for task 3A and a multi-layer perceptron for task 3B.

Team CIC 3A:11 experimented with a decision tree, a random forest, and a gradient boosting algorithms. They found the latter to perform best.

Team CIVIC-UPM [50] (3A:7 3B:8) participated in the two subtasks of task 3. They performed pre-processing, using a number of tools: (i) `ftfy` to repair Unicode and emoji errors, (ii) `ekphrasis` to perform lower-casing, normalizing percentages, time, dates, emails, phones, and numbers, (iii) `contractions` for abbreviation expansion, and (iv) NLTK for word tokenization, stop-words removal, punctuation removal and word lemmatization. Then, they combined `doc2vec` with transformer representations (Electra base, T5 small and T5 base, Longformer base, RoBERTa base and DistilRoBERTa base). They further used additional data from Kaggle's Ag News task, Kaggle's KDD2020, and Clickbait news detection competitions. Finally, they experimented with a number of classifiers such as Naïve Bayes, Random Forest, Logistic Regression with L1 and L2 regularization, Elastic Net, and SVMs. The best system for subtask 3A

used DistilRoBERTa-base on the text body with oversampling and a sliding window for dealing with long texts. Their best system for task 3B used RoBERTa-base on the title+body text with oversampling but no sliding window.

Team DLRG (3A: 3 3B: 4) experimented with a number of traditional approaches like Random Forest, Naïve Bayes and Logistic Regression as well as an online passive-aggressive classifier and different ensembles thereof. The best result was achieved by an ensemble of Naïve Bayes, Logistic Regression, and the Passive Aggressive classifier for task 3A. For task 3B, the Online Passive-Aggressive classifier outperformed all other approaches, including the considered ensembles.

Team GPLSI [56] (3A: 16) applied the RoBERTa transformer together with different manually-engineered features, such as the occurrence of dates and numbers or words from LIWC. Both the title and the body were concatenated as a single sequence of words. Rather than going for a single multiclass setting, they used two binary models considering the most frequent classes: false vs other and true vs other, followed by one three-class model.

Team MUCIC [58] (3A: 19 3B: 12) used a majority voting ensemble with three BERT variants. They applied BERT, Distilbert, and RoBERTa, and fine-tuned the pre-trained models.

Team NITK_NLP[48] (3A: 5 3B: 1) proposed an approach, that included pre-processing and tokenization of the news article, and then experimented with multiple transformer models. The final prediction was made by an ensemble.

Team NKovachevich [55] (3A: 13 3B: 3) created lexical features. They extracted the 500 most frequent word stems in the dataset and calculated the TF.IDF values, which they used in a multinomial Naïve Bayes classifier. A much better performance was achieved with an LSTM model that used GloVe embeddings. A little lower F1 value was achieved using BERT. They further found RoBERTa to perform worse than BERT.

Team NLP&IR@UNED [47] (3A: 4) experimented with four transformer architectures and input sizes of 150 and 200 words. In the preliminary tests, the best performance was achieved by ALBERT with 200 words. They also experimented with combining TF.IDF values from the text, all the features provided by the LIWC tool, and the TF.IDF values from the first 20 domain names returned by a query to a search engine. Unlike what was obtained in the dev dataset, the best results were obtained in the official competition with the approach based on TF.IDF, LIWC, and domain names.

Team NLytics (3A: 12 3B: 7) fine-tuned RoBERTa on the dataset for each of the sub-tasks. Since the data is unbalanced, they used under-sampling. They also truncated the documents to 512 words to fit into the RoBERTa input size.

Team NoFake (3A: 1 3B: 2) applied BERT without fine-tuning, but used an extensive amount of additional data for training, downloaded from various fact-checking websites.

Team Pathfinder [52] (3A: 9 3A: 10) participated in both tasks and used multinomial Naïve Bayes and random forest. The former performed better for both tasks. For task 3A, they merged the classed *false* and *partially false* into one class, which boosted the model performance by 41% (a non-official score mentioned in the paper).

Team Probity (3A: 20) addressed the multiclass fake news detection subtask, they used a simple LSTM architecture where they adopted word2vec embeddings to represent the news articles.

Team Qword [60] (3A: 23) applied pre-processing techniques, which included stop-word removal, punctuation removal and lemmatization using a Porter stemmer. The TF.IDF values

were calculated for the words. For these features, four classification algorithms were applied. The best result was given by Extreme Gradient Boosting.

Team SAUD (3A: 2) used an SVM with TF.IDF. They tried Logistic Regression, Multinomial Naïve Bayes, and Random Forest and found SVM to work best.

Team Sigmoid [57] (3A: 17) experimented with different traditional machine learning approaches, with multinomial Naïve Bayes performing best, and one deep learning approach, namely an LSTM with the Adam optimizer. The latter outperformed the more traditional approaches.

Team Spider (3A: 22) applies an LSTM, after a pre-processing consisting of stop-word removal and stemming.

Team UAICS [49] (3A: 6) experimented with various models including BERT, LSTM, Bi-LSTM, and feature-based models. Their submitted model is a Gradient Boosting with a weighted combination of three feature groups: bi-grams, POS tags, and lexical categories of words.

Team University of Regensburg [51] (3A: 8) used different fine-tuned variants of BERT with a linear layer on top and applied different approaches to address the maximum sequence length of BERT. Besides hierarchical transformer representations, they also experimented with different summarization techniques like extractive and abstractive summarization. They performed oversampling to address the class imbalance and extractive (using DistilBERT) and abstractive summarization (using distil-BART-CNN-12-6) before performing classification using fine-tuned BERT with a hierarchical transformer representation.

7. Conclusion and Future Work

We have presented a detailed overview of task 3, which focused on the classification of news articles with respect to the correctness of their main claims (task 3A) and their topical domain (task 3B). Most of the participants used transformer-based models like BERT and newer variants, RoBERTa being the most popular. For both subtask, the best run for each team outperformed the majority class baseline. Nonetheless, the F_1 -scores show large differences in the effectiveness of the applied approaches.

We plan a new iteration of task 3 in the CheckThat! lab augmenting the English data set and adding a multilingual setting.

Nevertheless, misinformation or Fake news will remain a social issue that cannot be purely solved by technological advancement. Also, education toward information literacy is an important strategy to defend against misinformation online, and offline [72].

8. Acknowledgements

We are thankful for the CheckThat! organizers for supporting this task. We also thank the student volunteers for helping with the annotation of the data.

References

- [1] D. Bawden, L. Robinson, The dark side of information: overload, anxiety and other paradoxes and pathologies, *J. Inf. Sci.* 35 (2009) 180–191. URL: <https://doi.org/10.1177/0165551508095781>. doi:10.1177/0165551508095781.
- [2] R. Zafarani, X. Zhou, K. Shu, H. Liu, Fake news research: Theories, detection strategies, and open problems, in: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, ACM, 2019, pp. 3207–3208. URL: <https://doi.org/10.1145/3292500.3332287>. doi:10.1145/3292500.3332287.
- [3] 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, *LNCS (12880)*, Springer, 2021.
- [4] P. Nakov, G. D. S. Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, The CLEF-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: D. Hiemstra, M. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II, volume 12657 of Lecture Notes in Computer Science*, Springer, 2021, pp. 639–649. URL: https://doi.org/10.1007/978-3-030-72240-1_75. doi:10.1007/978-3-030-72240-1_75.
- [5] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, *ACM Comput. Surv.* 53 (2020) 109:1–109:40. URL: <https://doi.org/10.1145/3395046>. doi:10.1145/3395046.
- [6] X. Zhang, A. A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion, *Inf. Process. Manag.* 57 (2020) 102025. URL: <https://doi.org/10.1016/j.ipm.2019.03.004>. doi:10.1016/j.ipm.2019.03.004.
- [7] M. Hardalov, A. Arora, P. Nakov, I. Augenstein, A survey on stance detection for mis- and disinformation identification, *CoRR abs/2103.00242 (2021)*. URL: <https://arxiv.org/abs/2103.00242>. arXiv:2103.00242.
- [8] S. I. Manzoor, J. Singla, et al., Fake news detection using machine learning approaches: A systematic review, in: *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, IEEE, 2019, pp. 230–234. doi:<https://doi.org/10.1109/ICOEI.2019.8862770>.
- [9] X. Li, X. L. Dong, K. Lyons, W. Meng, D. Srivastava, Truth finding on the deep web, *Proceedings of the VLDB Endowment* 6 (2012) 97–108. doi:10.14778/2535568.2448943.
- [10] X. Li, W. Meng, C. Yu, T-verifier: Verifying truthfulness of fact statements, in: *2011 IEEE 27th International Conference on Data Engineering*, IEEE, 2011, pp. 63–74.
- [11] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, J. Han, A survey on truth discovery, *ACM Sigkdd Explorations Newsletter* 17 (2016) 1–16.
- [12] K. Popat, S. Mukherjee, J. Strötgen, G. Weikum, Credibility assessment of textual claims on the web, in: *Proceedings of the 25th ACM International on Conference on Information*

and Knowledge Management, 2016, pp. 2173–2178.

- [13] G. K. Shahi, D. Nandini, FakeCovid – a multilingual cross-domain fact check news dataset for covid-19, in: Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media, 2020. URL: http://workshop-proceedings.icwsm.org/pdf/2020_14.pdf.
- [14] S. Helmstetter, H. Paulheim, Weakly supervised learning for fake news detection on twitter, in: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2018, pp. 274–277.
- [15] D. Röchert, G. K. Shahi, G. Neubaum, S. Ross, Björn Stieglitz, The networked context of covid-19 misinformation: informational homogeneity on youtube at the beginning of the pandemic, arXiv preprint (2021).
- [16] M. L. Ba, L. Berti-Equille, K. Shah, H. M. Hammady, VERA: A platform for veracity estimation over web data, in: Proceedings of the 25th International Conference on World Wide Web, WWW '16, 2016, pp. 159–162.
- [17] R. Baly, G. Karadzhov, J. An, H. Kwak, Y. Dinkov, A. Ali, J. Glass, P. Nakov, What was written vs. who read it: News media profiling using text analysis and social media context, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL '20, 2020, pp. 3364–3374.
- [18] G. Karadzhov, P. Nakov, L. Márquez, A. Barrón-Cedeño, I. Koychev, Fully automated fact checking using external sources, in: Proceedings of RANLP 2017, 2017, pp. 344–353.
- [19] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, M. Cha, Detecting rumors from microblogs with recurrent neural networks, in: Proceedings of IJCAI, 2016.
- [20] S. Mukherjee, G. Weikum, Leveraging joint interactions for credibility analysis in news communities, in: Proceedings of CIKM' 15, 2015, pp. 353–362.
- [21] V.-H. Nguyen, K. Sugiyama, P. Nakov, M.-Y. Kan, FANG: Leveraging social context for fake news detection using graph representation, in: Proceedings of the 29th ACM International Conference on Information Knowledge Management, CIKM '20, 2020, p. 1165–1174.
- [22] A. Zubiaga, M. Liakata, R. Procter, G. W. S. Hoi, P. Tolmie, Analysing how people orient to and spread rumours in social media by looking at conversational threads, PLoS ONE 11 (2016).
- [23] 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.
- [24] 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.
- [25] 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.
- [26] Z. Zhao, P. Resnick, Q. Mei, Enquiring minds: Early detection of rumors in social media from enquiry posts, in: Proceedings of the 24th International Conference on World Wide Web, WWW'15, 2015, pp. 1395–1405.
- [27] R. Oshikawa, J. Qian, W. Y. Wang, A survey on natural language processing for fake news detection, in: Proceedings of the 12th Language Resources and Evaluation Conference, LREC '20, 2020, pp. 6086–6093.
- [28] M. Hardalov, I. Koychev, P. Nakov, In search of credible news, in: C. Dichev, G. Agre

- (Eds.), *Artificial Intelligence: Methodology, Systems, and Applications - 17th International Conference, AIMS A 2016, Varna, Bulgaria, September 7-10, 2016, Proceedings*, volume 9883 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 172–180. URL: https://doi.org/10.1007/978-3-319-44748-3_17. doi:10.1007/978-3-319-44748-3_17.
- [29] G. K. Shahi, A. Dirkson, T. A. Majchrzak, An exploratory study of covid-19 misinformation on twitter, *Online social networks and media* (2021) 100104. doi:10.1016/j.osnem.2020.100104.
- [30] G. K. Shahi, T. A. Majchrzak, Exploring the spread of covid-19 misinformation on twitter, EasyChair Preprint no. 6009, EasyChair, 2021.
- [31] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, A. Zubiaga, SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours, in: *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval '17, 2017*, pp. 69–76.
- [32] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, L. Derczynski, SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours, in: *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval '19, 2019*, pp. 845–854.
- [33] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, C. Cherry, SemEval-2016 task 6: Detecting stance in tweets, in: *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16, 2016*, pp. 31–41.
- [34] G. Da San Martino, A. Barrón-Cedeno, H. Wachsmuth, R. Petrov, P. Nakov, SemEval-2020 task 11: Detection of propaganda techniques in news articles, in: *Proceedings of the 14th Workshop on Semantic Evaluation, SemEval '20, 2020*, pp. 1377–1414.
- [35] T. Mihaylova, G. Karadzhov, P. Atanasova, R. Baly, M. Mohtarami, P. Nakov, SemEval-2019 task 8: Fact checking in community question answering forums, in: *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval '19, 2019*, pp. 860–869.
- [36] K. Pogorelov, D. T. Schroeder, L. Burchard, J. Moe, S. Brenner, P. Filkukova, J. Langguth, FakeNews: Corona virus and 5G conspiracy task at MediaEval 2020, in: *MediaEval 2020 Workshop*, 2020.
- [37] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL '18, 2018*, pp. 809–819.
- [38] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fact extraction and verification (FEVER) shared task, *CoRR abs/1811.10971* (2018). URL: <http://arxiv.org/abs/1811.10971>. arXiv:1811.10971.
- [39] A. Chernyavskiy, D. Ilvovsky, P. Nakov, Whatthewikifact: Fact-checking claims against wikipedia, *CoRR abs/2105.00826* (2021). URL: <https://arxiv.org/abs/2105.00826>. arXiv:2105.00826.
- [40] K. Nakamura, S. Levy, W. Y. Wang, r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection, *arXiv preprint arXiv:1911.03854* (2019).
- [41] A. Kazemi, K. Garimella, G. K. Shahi, D. Gaffney, S. A. Hale, Tiplines to combat misinformation on encrypted platforms: A case study of the 2019 indian election on whatsapp, *CoRR abs/2106.04726* (2021). URL: <https://arxiv.org/abs/2106.04726>. arXiv:2106.04726.

- [42] G. K. Shahi, T. A. Majchrzak, Exploring the spread of covid-19 misinformation on twitter, arXiv preprint (2021).
- [43] G. K. Shahi, AMUSED: An annotation framework of multi-modal social media data, 2020. arXiv:2010.00502.
- [44] G. K. Shahi, A multilingual domain identification using fact-checked articles: A case study on covid-19 misinformation, arXiv preprint (2021).
- [45] G. K. Shahi, J. M. Struß, T. Mandl, CT-FAN-21 corpus: A dataset for Fake News Detection, 2021. URL: <https://doi.org/10.5281/zenodo.4714517>. doi:10.5281/zenodo.4714517.
- [46] S. Kumari, NoFake at CheckThat! 2021: Fake news detection using BERT, arXiv preprint (2021).
- [47] 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: [73], 2021.
- [48] H. R. L. A. M., NITK_NLP at CLEF CheckThat! 2021: Ensemble transformer model for fake news classification, in: [73], 2021.
- [49] C.-G. Cusmuluc, M.-A. Amarandei, I. Pelin, V.-I. Cociorva, A. Iftene, UAICS at CheckThat! 2021: Fake news detection, in: [73], 2021.
- [50] Álvaro Huertas-Garcia, J. Huertas-Tato, A. Martín, D. Camacho, CIVIC-UPM at CheckThat! 2021: Integration of transformers in misinformation detection and topic classification, in: [73], 2021.
- [51] P. Hartl, U. Kruschwitz, University of Regensburg at CheckThat! 2021: Exploring text summarization for fake news detection, in: [73], 2021.
- [52] W. K. Tsoplefack, Classifier for fake news detection and topical domain of news articles, in: [73], 2021.
- [53] N. Ashraf, S. Butt, G. Sidorov, A. Gelbukh, Cic at CheckThat! 2021: Fake news detection using machine learning and data augmentation, in: [73], 2021.
- [54] S. Sohan, H. S. Rajon, A. Khusbu, M. S. Islam, M. A. Hasan, Black Ops at CheckThat! 2021: User profiles analyze of intelligent detection on fake tweets notebook in shared task, in: [73], 2021.
- [55] N. Kovachevich, Nkovachevich at CheckThat! 2021: Bert fine-tuning approach to fake news detection, in: [73], 2021.
- [56] R. Sepúlveda-Torres, E. Saquete, GPLSI team at CLEF CheckThat! 2021: Fine-tuning BETO and RoBERTa, in: [73], 2021.
- [57] A. A. M. Sardar, S. A. Salma, M. S. Islam, M. A. Hasan, T. Bhuiyan, Team Sigmoid at CheckThat! 2021: Multiclass fake news detection with machine learning, in: [73], 2021.
- [58] F. Balouchzahi, H. Shashirekha, G. Sidorov, MUCIC at CheckThat! 2021: FaDo-fake news detection and domain identification using transformers ensembling, in: [73], 2021.
- [59] S. S. Ashik, A. R. Apu, N. J. Marjana, M. A. Hasan, M. S. Islam, M82B at CheckThat! 2021: Multiclass fake news detection using BiLSTM, in: [73], 2021.
- [60] R. S. Utsha, M. Keya, M. A. Hasan, M. S. Islam, Qword at CheckThat! 2021: An extreme gradient boosting approach for multiclass fake news detection, in: [73], 2021.
- [61] L. Dumani, P. J. Neumann, R. Schenkel, A framework for argument retrieval - ranking argument clusters by frequency and specificity, in: *Advances in Information Retrieval - 42nd European Conference on IR Research (ECIR)*, volume 12035 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 431–445.

- [62] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in indo-european languages, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 87–111. URL: <http://ceur-ws.org/Vol-2826/T2-1.pdf>.
- [63] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018, in: P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J. C. de Albornoz (Eds.), Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018, volume 2150 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 214–228. URL: <http://ceur-ws.org/Vol-2150/overview-AMI.pdf>.
- [64] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, E. Sherly, J. P. McCrae, Overview of the track on sentiment analysis for dravidian languages in code-mixed text, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 480–489. URL: <http://ceur-ws.org/Vol-2826/T4-1.pdf>.
- [65] A. Tchechmedjiev, P. Fafalios, K. Boland, M. Gasquet, M. Zloch, B. Zapilko, S. Dietze, K. Todorov, Claimskg: A knowledge graph of fact-checked claims, in: C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. F. Cruz, A. Hogan, J. Song, M. Lefrançois, F. Gandon (Eds.), The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II, volume 11779 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 309–324. URL: https://doi.org/10.1007/978-3-030-30796-7_20. doi:10.1007/978-3-030-30796-7_20.
- [66] A. Pritzkau, NLytics at CheckThat! 2021: Multi-class fake news detection of news articles and domain identification with RoBERTa - a baseline model, in: [73], 2021.
- [67] F. Sakketou, N. Ampazis, A constrained optimization algorithm for learning glove embeddings with semantic lexicons, *Knowl. Based Syst.* 195 (2020) 105628. URL: <https://doi.org/10.1016/j.knosys.2020.105628>. doi:10.1016/j.knosys.2020.105628.
- [68] B. Majumdar, M. R. Bhuiyan, M. A. Hasan, M. S. Islam, S. R. Haider Noori, Probity at CheckThat! 2021: Multi class fake news detection using LSTM approach, in: [73], 2021.
- [69] G. K. Shahi, I. Bilbao, E. Capecci, D. Nandini, M. Choukri, N. Kasabov, Analysis, classification and marker discovery of gene expression data with evolving spiking neural networks, in: International Conference on Neural Information Processing, Springer, 2018, pp. 517–527.
- [70] D. Nandini, E. Capecci, L. Koefoed, I. Laña, G. K. Shahi, N. K. Kasabov, Modelling and analysis of temporal gene expression data using spiking neural networks, in: Neural Information Processing - 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13-16, 2018, Proceedings, Part I, volume 11301 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 571–581. URL: https://doi.org/10.1007/978-3-030-04167-0_52. doi:10.1007/978-3-030-04167-0_52.
- [71] G. K. Shahi, D. Nandini, S. Kumari, Inducing schema. org markup from natural language

context, *Kalpa Publications in Computing* 10 (2019) 38–42.

- [72] S. Dreisiebner, A. K. Polzer, L. Robinson, P. Libbrecht, J. Boté-Vericad, C. Urbano, T. Mandl, P. Vilar, M. Zumer, M. Juric, F. Pehar, I. Stricevic, Facilitation of information literacy through a multilingual MOOC considering cultural aspects, *J. Documentation* 77 (2021) 777–797. URL: <https://doi.org/10.1108/JD-06-2020-0099>. doi:10.1108/JD-06-2020-0099.
- [73] 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, CEUR-WS.org, 2021.