Overview of the CLEF-2022 CheckThat! Lab: Task 3 on Fake News Detection

Juliane Köhler[1], Gautam Kishore Shahi[2], Julia Maria Struß[1], Michael Wiegand[3], Melanie Siegel[4], Thomas Mandl[5] and Mina Schütz[6]

[1]University of Applied Sciences Potsdam, Germany
[2]University of Duisburg-Essen, Germany
[3]Alpen-Adria-Universität Klagenfurt, Austria
[4]Darmstadt University of Applied Sciences, Germany
[5]University of Hildesheim, Germany
[6]AIT Austrian Institute of Technology GmbH, Austria

Abstract

This paper describes the results of the CheckThat! Lab 2022 Task 3. This is the fifth edition of the lab, which concentrates on the evaluation of technologies supporting three tasks related to factuality. Task 3 is designed as a multi-class classification problem and focuses on the veracity of German and English news articles. The German subtask is ought to be solved using an cross-lingual approach while the English subtask was offered as mono-lingual task. The participants of the lab were provided an English training, development and test dataset as well as a German test dataset. In total, 25 teams submitted successful runs for the English subtask and 8 for the German subtask. The best performing system for the mono-lingual subtask achieved a macro F1-score of 0.339. The best system for the cross-lingual task achieved a macro F1-score of 0.242. In the paper at hand we will elaborate on the process of data collection, the task setup, the evaluation results and give a brief overview of the participating systems.

Keywords

Misinformation, Fake News, Text Classification, Evaluation, Deep Learning, Machine Learning

1. Introduction

During the COVID-19 pandemic, the World Health Organization did not only speak of a pandemic, but even an Infodemic, due to the vast amount of false information about the disease[1]. Regarding the huge societal impact of misinformation, research on the topic received a lot of attention in the past years. To counter the spread of wrong and potentially harmful information, researchers explored different approaches to detect fake news in different media forms such as social media [1, 2], news papers [3, 4], deep fakes [5, 6] and others [7, 8]. The CheckThat! Lab at CLEF contributes to those research efforts by offering tasks along the

CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5-8, 2022, Bologna, Italy

juliane.koehler@fh-potsdam.de (J. Köhler); gautam.shahi@uni-due.de (G. K. Shahi); struss@fh-potsdam.de (J. M. Struß); michael.wiegand@aau.at (M. Wiegand); melanie.siegel@h-da.de (M. Siegel); mandl@uni-hildesheim.de (T. Mandl); mina.schuetz@ait.ac.at.at (M. Schütz)
0000-0002-7175-5895 (J. Köhler); 0000-0001-6168-0132 (G. K. Shahi); 0000-0001-9133-4978 (J. M. Struß); 0000-0002-5403-1078 (M. Wiegand); 0000-0002-8398-9699 (T. Mandl)

[1] https://www.who.int/health-topics/infodemic#tab=tab_1

full verification pipeline with high quality data and corresponding evaluation environments. Thus, fostering the development of approaches to perform fake news identification and provide tools to support individuals. This lab offers three tasks [9, 10] which are all described in the lab's overview paper [11]. The paper at hand provides a detailed overview of Task 3 of the `CheckThat!` Lab 2022. The task focuses on predicting the truthfulness of articles and is further elaborated in section 3. Furthermore, it addresses the challenge of providing a dataset of genuine information and misinformation in news articles. While users might be careful to trust low-quality information on social media applications, false information in news articles poses a threat, as victims might be less suspicious. Accordingly, efforts must be made to create high-quality datasets that are needed for the conduction of fruitful experiments.

The remainder of this paper is organized as follows: Section 2 gives an overview on the state-of-the-art research, section 3 provides detailed information on the task, while section 4 explains the process of data collection. Section 5 presents the evaluation results and gives an overview of the participants approaches. Section 6 describes our baseline classifier and lists detailed descriptions of the different approaches used the by the individual participating teams. Finally, we provide a brief conclusion and an outlook on potential future work in section 7.

2. Related Work

Much work was recently dedicated towards the identification of misinformation in general [12, 13, 14] and in particular in social media [15, 16]. Fake news detection for social media poses several challenges which require more research. Among them are visual content [8] and fast dissemination [1, 2]. News articles can be considered as less complex than social media. Nevertheless, current detection systems can still not provide satisfying results as the Task 3a of `CheckThat!` 2021 showed [17]. Furthermore, most studies only model fake news detection as a binary classification problem [1, 2, 3, 4, 14, 18, 19, 20]. Task 3 of the `CheckThat!` 2022 Lab therefore offers a task on multi-class classification of news articles.

Several other initiatives related to the `CheckThat!` Lab at CLEF aim to advance research on fake news detection. MediaEval 2021 offered a text-based fake news and conspiracy theory detection task with three subtasks [21]. Their data originates from Twitter posts and news articles and the main topic in the data is COVID-19. A similar task was already hosted by MediaEval 2020 [22] focusing on COVID-19 and 5G conspiracy theories.

RumourEval [23, 24], as part of SemEval, addressed stance detection and also classifying tweets according to the truthfulness. Other SemEval tasks concerned stance [25], and propaganda detection [26] as well as fact-checking in community question answering forums [27]. FEVER [28, 29] focused on Wikipedia data for supporting or invalidating claims.

In 2019, the Qatar International Fake News Detection and Annotation Contest [2] was conducted. The task description is similar to last year's iteration of `CheckThat!`'s Task 3 [17]. The first subtask focused on the classification of news articles, detecting if an article is fake or legitimate [30]. The second subtask was on deciding on the topical domain of a news article and the third addressed the automatic distinction of human and bot accounts on Twitter.

[2] https://sites.google.com/view/fakenews-contest

Another related shared task is the Fake News Challenge Stage 1 (FNC-I)[3], which centered around stance detection. The aim was to develop automatic systems that, given a random pair of a title and the body of two different or the same article, classify the pair in one of four stance classes: *Agrees* (if the body text agrees with the title), *Disagrees* (if the body text disagrees with the title), *Discusses* (if the body text discusses the title without taking a position), *Unrelated* (if the body text and title are unrelated). Participants were given a dataset that consists of titles and bodies of news articles as well as a the stance dataset [4]. In the latter, a pair of title and body were allocated the according stance. The most successful submission[5] applied both a XGBoost classifier and a 1D convolutional neural network classifier. The weighted average of those two classifiers was taken as output.

The technology for detecting misinformation can be broadly categorised into knowledge based approaches which use knowledge bases and compare claims to them in some way (e.g. [31]) and text classification approaches which learn to distinguish between texts with wrong information and texts with correct information based on examples (e.g. [32]). Task 3 of `CheckThat!` 2022 is dedicated to evaluate text classification methods.

3. Task Description

The `CheckThat!` Task 3 evaluates systems which predict the veracity of news articles and is designed as a multi-class classification problem. In 2022, the second iteration of the task was conducted. As in 2021, the task is offered as a monolingual task in English. Additionally – in line with the general CLEF mission – the task was also offered as a cross-lingual task this year, providing English training and German test data. The overall problem definition is equivalent to Subtask 3A from last year's task:

Task 3: Multi-class fake news detection of news articles. Given the text and the 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. [33, 34] and the definitions for the four categories are 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.

[3] http://www.fakenewschallenge.org/

[4] https://github.com/FakeNewsChallenge/fnc-1

[5] https://github.com/Cisco-Talos/fnc-1

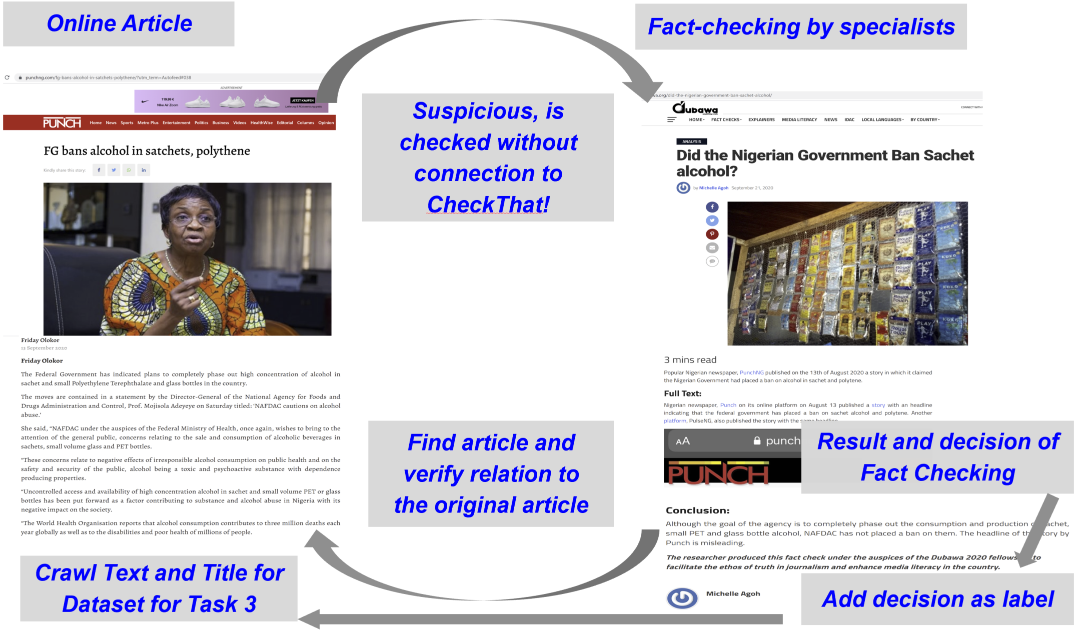
Online Article
Fact-checking by specialists
Suspicious, is checked without connection to CheckThat!
FG bans alcohol in satchets, polythene
Did the Nigerian Government Ban Sachet alcohol?
Find article and verify relation to the original article
Result and decision of Fact Checking
Crawl Text and Title for Dataset for Task 3
Add decision as label

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

4. Data Description

For this task, we aimed for two high-quality, real-life fake news datasets that address a wide range of topics in two languages: English and German. Fact-checking claims is not only a time-consuming task, but also requires training and experience. Therefore, to ensure high data quality, we relied on expert evaluations of claims in news articles that were documented on fact-checking services' websites. In the following section, the process of data collection and the dataset itself will be described in detail. A summary of the approach is depicted in Figure 1.

4.1. Crawling Fact-Checking Reports

The general procedure for data crawling was adopted from last year's iteration of the task [17]: First, fact-checking sites with an appropriate structure for crawling and focus had to be found. For the English data, the same fact-checking sites were used as last year. For the German data, an analysis of available fact-checking sites was conducted, and seven websites were judged as suitable for crawling regarding the purposes of this task. The crawling process was based on the AMUSED framework [35]. From each fact-checking site, we collected the report about a claim, the experts' judgement on the truthfulness of the claim, links to the potential source of a claim, and information on the type of the source (e. g. news article, social media post, video). Based on the source type, automatic filtering was applied, removing all social media posts and non-textual documents.

If available, the metadata was accessed via the JSON format using the *ClaimReview*-type defined by Schema.org. However, many websites did not make use of this format and a clear general position of a source link did not exist. Therefore, the first three links of a report were collected, based on the observation that one of those usually referred to the claim source. A manual check of the collected links given in the reports followed to identify the correct source link. For the German data around 1,300 links from roughly 650 reports were manually examined this way. For the English data the same was done for around 1,100 links from roughly 780 reports.

Furthermore, to generate more articles for the category *true*, we made use of sources the fact-checking experts referred to, in order to validate and proof their judgements. This decision was made, because those references were implicitly judged as reliable. On top of that, these articles covered similar topics as the original claim, thus counteracting a topical bias between classes. The collection of the according URLs was done manually as well.

After the collection and evaluation of links, roughly 1,500 articles (779 German, 711 English) remained for scraping.

4.2. Scraping Articles from the Web

From the remaining article candidates and their corresponding links; title and text were extracted in an automatic scraping process. Due to very diverse websites, the creation of tailored scrapers was not feasible. Instead, the h1-tags were extracted as titles and the contents of the p-tags as text, excluding footer contents. For data with missing titles and texts or with a text length below a threshold of 500 characters, the according articles were checked manually again. Often, those articles were not extracted correctly, sitting behind a paywall, or having already removed for the public. If possible, the missing values were added to the data manually or otherwise the article was deleted. Since many fact checking sites relied on archived versions of an article, different URLs sometimes lead to the same content. Those duplicates were also removed from the corpus.

4.3. Data Set for Task 3

In total, we relied on data from 20 different websites with AFP being used for both languages). Each fact-checking agency made use of their customized labels (see Table 1 for examples). Thus, instead of providing the original labels, we merged the labels with a similar meaning to one of our four categories: *false*, *partially false*, *true* or *other*.

The participants of Task 3 were provided an English training set that consisted of last year's training set (900 articles) and an English development set that served as test set in the last iteration (364 articles). The newly collected data was given as test set without labels to the participants. The test sets consist of the title and text from 612 English and 586 German articles. An overview about the distribution of the different classes in the fake news detection corpus, CT-FAN-22 [36], is given in Table 3. Each dataset included a unique identifier that was given to each individual article, the title of an article, the main text of an article and the respective class label. Table 2 shows some sample data.

Table 1
Examples for labels merged to build the final set of classes for task 3

task label	original label English	original label German
false	fake, false, inaccurate, inaccurate with consideration, incorrect, not true, pants-fire	Falsch, Stimmt nicht, Fälschung, Frei erfunden, Manipuliert, Fälschung
partially false	half-true, imprecise, mixed, partially true, partly true, barely-true, misleading	Irreführend, Teilweise falsch, Größtenteils falsch, Gößtenteils richtig, Stimmt eher nicht, Halbwahrheit
true	accurate, correct, true	Wahr, Stimmt, Richtig
other	debated, other, unclear, unknown, unsupported	Keine Beweise, Keine Belege, Keine Hinweise, Unbelegt

Table 2
Sample data for task 3

public_id	title	text	rating
8666812565752130848 7648037552383161291	Geheime Planspiele von „Grünen" und SPD-Linken: Esken soll an Stelle von Scholz Kanzlerin werden!	Es ist der große Plan B, über den vor der Wahl nichts nach draußen dringen darf: Auch wenn SPD-Kanzlerkandidat Olaf Scholz am Sonntag als Sieger aus der Bundestagswahl hervorgehen sollte, könnten „Grüne" und SPD-Linke ihn auf dem Weg ins Kanzleramt noch zu Fall bringen. [...][6]	False
27007788890674305467 3881300019061126049	Quebec's liquor and cannabis stores will require vaccine passport as of Jan. 18	Health Minister Christian Dubé hopes the measure encourages more Quebecers to get vaccinated. Quebecers will have to present proof of vaccination to access the province's liquor and cannabis stores as of Tuesday, Jan. 18. [...][7]	True

5. Submissions and Results

In this section, we present an overview of all submissions for Task 3 of the CheckThat! lab 2022. Each team could submit up to 200 runs. Yet, only the last submission was taken into account for the evaluation. In total, there were 32 submissions evaluated for detecting English fake news and

[6] https://deutschlandkurier.de/2021/09/geheime-planspiele-von-gruenen-und-spd-linken-esken-soll-an-stelle-von-scholz-kanzlerin-werden/

[7] https://montrealgazette.com/news/local-news/quebecs-liquor-and-cannabis-stores-will-require-vaccine-passport-as-of-jan-18

Table 3
The number of documents and class distribution for the CT-FAN-22 corpus for English (left) and German (right) fake news detection

Class	Training	Development	Test
False	465	113	315
True	142	69	210
Partially False	217	141	56
Other	76	41	31
Total	900	364	612

Class	Test
False	191
True	243
Partially False	97
Other	55
Total	586

16 submissions for the German part of the task. Out of the 32 submissions for English detection, 7 were rejected due to the submission being either incorrectly formatted or incomplete. The same is true for half, that is, 8 of the submissions for the German detection. However, 6 out of those 8 flawed submissions were rejected, because they contained classification results for the English instead of the German test data.

Out of 26 teams that successfully submitted a solution that fully complied with our format specifications for at least one of the two subtasks, most teams only participated in the English monolingual subtask (25 accepted submissions). Yet, 8 teams also successfully submitted runs to the cross-lingual subtask.

Both subtasks are classification tasks. Therefore, we used accuracy and the macro-F_1 score for evaluation and ranked the systems by the latter.

Table 4 provides an overall summary of system performances in terms of macro-F_1 score. In both subtasks the best system score is still fairly low. This underlines the general difficulty of both tasks. It comes as no surprise that the best score in the cross-lingual task is below that of the monolingual task, since the latter is a more difficult problem.

Table 4 shows that the baseline classifier, further described in Section 6.1, is very strong for both subtasks, this is particularly true for the monolingual task. In both cases, it is notably above the median. The median in the monolingual task is a bit closer to the best score than in the cross-lingual task. This suggests that the number of stronger systems is greater in the monolingual task.

In both tasks, the range between worst and best score is still not insignificant. Despite the fairly similar system design of the different participants within the same subtask as outlined in section 6 (i.e. fine-tuning a transformer), there still seem to be several degrees of freedom which have a major impact on overall performance (e.g. hyperparameter settings or the particular choice of the language model).

In Tables 5 and 6 the latest submission of each team for both subtasks is given.

5.1. Results of fake news categorization of news articles for English data

In total, 25 teams attempted to solve the first task, which was the the monolingual English task. The best system for the monolingual subtask was submitted by team **iCompass** [37] (macro-

Table 4
Summary statistics for overall macro F1-scores in the two subtasks.

Subtask	#Teams	Baseline	Min	Max	Median
Monolingual	25	0.312	0.117	0.339	0.271
Cross-lingual	8	0.242	0.111	0.290	0.191

averaged F1 score: 0.339). They applied $bert\text{-}base\text{-}uncased$ on title and main text seperately and concatenated the results. The model was fine-tuned on the task-specific training data. They also experimented with RoBERTa for which they got worse results. No additional external resources were employed in the final classifier.

The second-best system in this subtask was submitted by team **NLP&IR@UNED** [38] (macro-averaged F1 score: 0.332). The team made use of an ensemble classifier. It was built out of a Funnel Transformer and a Feed Forward Neural Network. The features were extracted by the $LIWC$ text analysis tool.

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

5.2. Results of fake news categorization of news articles for German data

In total 8 teams attempted to solve the second subtask, which was the English-German cross-lingual task. Team **ur-iw-hnt** [39], as the team with the most successful submission (macro-averaged F1 score: 0.290), translated the first 5000 tokens of an article from the German test data using the service of Google Translate. They applied an extractive summarization technique and a $BERT_{Large}$ model for the multi-class classification.

Team **NITK-IT_NLP** [40], which was the team with the second-best submission, divided the text of the news articles into windows of 500 tokens. Those windows are shifted over the text in order not to lose context. They experimented with different transformer models, with a $mDeBERTa$ model yielding the best results. The individual results of all 8 submissions are depicted in Table 6.

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

6. Discussion of the Approaches Used

Before we give a summary of the different classification approaches in section 6.2, we will first describe the baseline classifier that was used in this year's shared task (section 6.1).

6.1. The Baseline Classifier

To have a starting point for the participants, we created a baseline system. The model used for the CheckThat! 2022 Task 3 baseline is a standard bert-base-cased model from HuggingFace (no lower-casing during training). The downloaded pre-trained model is originally trained on English data and was fine-tuned on the 900 articles from the CheckThat! training set. The

Table 5
English: Official evaluation results for English Fake News Detection ranked by the macro-F_1 score, including the F_1 scores for individual classes and the overall accuracy

	Team	True	False	Partially False	Other	Accuracy	Macro-F1
1	iCompass [37]	0.383	0.721	0.173	0.080	0.547	0.339
2	NLP&IR@UNED [38]	0.446	0.729	0.097	0.057	0.541	0.332
3	Awakened [41]	0.328	0.744	0.185	0.035	0.531	0.323
4	UNED	0.346	0.725	0.191	0.000	0.544	0.315
Baseline		0.244	0.701	0.157	0.144	0.480	0.312
5	NLytics [42]	0.339	0.707	0.184	0.000	0.513	0.308
6	SCUoL [43]	0.377	0.709	0.133	0.000	0.526	0.305
7	NITK-IT_NLP [40]	0.325	0.734	0.133	0.000	0.536	0.298
8	CIC [44]	0.111	0.682	0.215	0.136	0.475	0.286
9	ur-iw-hnt [39]	0.290	0.733	0.110	0.000	0.533	0.283
10	BUM [45]	0.207	0.694	0.140	0.063	0.472	0.276
11	boby232	0.255	0.676	0.126	0.045	0.475	0.275
12	HBDCI [46]	0.177	0.708	0.209	0.000	0.508	0.273
13	DIU_SpeedOut	0.195	0.706	0.182	0.000	0.521	0.271
14	DIU_Carbine	0.192	0.626	0.157	0.056	0.472	0.258
15	CODE [47]	0.126	0.662	0.203	0.029	0.444	0.255
16	MNB	0.160	0.701	0.142	0.000	0.507	0.251
17	subMNB	0.160	0.701	0.142	0.000	0.507	0.251
18	FoSIL [48]	0.141	0.670	0.169	0.022	0.462	0.251
19	TextMinor [49]	0.250	0.555	0.086	0.048	0.377	0.235
20	DLRG	0.009	0.694	0.092	0.000	0.513	0.199
21	DIU_Phoenix	0.420	0.040	0.092	0.000	0.278	0.159
22	AIT_FHSTP [50]	0.280	0.146	0.154	0.039	0.199	0.155
23	DIU_SilentKillers	0.407	0.070	0.135	0.000	0.260	0.153
24	DIU_Fire71	0.430	0.006	0.094	0.000	0.275	0.133
25	AI Rational	0.296	0.000	0.196	0.090	0.098	0.117

training parameters were: a batch size of 8, a maximum sequence length of 512, 10 epochs and a learning rate of 3e-5.

Since only one article was longer than the maximum sequence length, we did not use passage classification or windows for training. We adopted AdamW as an optimizer with a linear scheduler without warm up. For the training/validation split we used 90% of the training set for training and 10% for validation. The training loss was 0.04 after 10 epochs. The outputs on the validation set (training data) after completion of training were the following:

- Accuracy: 0.56
- Precision: 0.44 (macro-averaged)
- Recall: 0.44 (macro-averaged)

Table 6
German: Official evaluation results for German Fake News Detection ranked by the macro-F_1 score, including the F_1 scores for individual classes and the overall accuracy

	Team	True	False	Partially False	Other	Accuracy	Macro-F1
1	ur-iw-hnt [39]	0.401	0.536	0.189	0.033	0.427	0.290
	Baseline	0.405	0.328	0.029	0.204	0.280	0.242
2	NITK-IT_NLP [40]	0.268	0.490	0.077	0.063	0.362	0.225
3	UNED	0.298	0.166	0.210	0.162	0.213	0.209
4	AIT_FHSTP [50]	0.378	0.168	0.151	0.081	0.254	0.195
5	Awakened [41]	0.098	0.452	0.194	0.000	0.283	0.186
6	CIC [44]	0.000	0.449	0.240	0.000	0.282	0.172
7	NoFake	0.000	0.492	0.000	0.000	0.326	0.123
8	AI Rational	0.268	0.000	0.166	0.122	0.114	0.111

- F1: 0.42 (macro-averaged)

The results on the test set show that there was a problem with overfitting (see Table 5).

The baseline model is trained on the title and text content of the articles. It is based on former work in [51], where also a bert-base-cased was used on another fake news detection dataset. In that work, it was shown that using the titles in front of the body content boosted the accuracy. This boost could also be observed on the CheckThat! Task 3 data. For the German data, we used automatic translation of German texts to English and then applied the baseline model. The results of the baseline system for German can be found in Table 6.

6.2. Classification Approaches

Most experiments involved deep learning models (16 teams), especially applications of BERT (12 teams), RoBERTa (6 teams) or other BERT versions (in total 8 teams) were popular. However, almost as many teams (14 teams) experimented with feature-based supervised-learning approaches as well. Examples are SVMs (10 teams), Logistic Regression (9 teams), Random Forests (8 teams) and Naive Bayes (7 teams). Yet, the majority merely fine-tuned a pre-trained language model and only very few experimented with other approaches.

Although still quite many participants also experimented with feature-based approaches, only very few teams incorporated a more non-standard feature design, i.e. features other than bag of words, n-grams or word embeddings. One team (**NLP&IR@UNED** [38]) employed features from LIWC [52] , one other team (**HBDCI** [46]) made use of surface features to capture misspellings and repeated sentences. One further team (**FoSIL** [48]) implemented a special feature selection scheme using *human behavior based optimization.*

Surprisingly, only two participants (**BUM** [45] and **NLP&IR@UNED** [38]) considered employing an ensemble of different classifiers, despite the fact that this procedure is a simple and established method for effectively combining individual classifiers of varying performances.

Only very few teams used additional processing techniques which are not part of standard text classification algorithms. **BUM** [45] exploited Wikipedia for evidence retrieval. Team **ur-iw-hnt** [39] incorporate summarization techniques (both abstractive and extractive ones) in order to account suitable for long documents.

In order to bridge the language gap between the English training data and German test data in the cross-lingual subtask, either a multi-lingual language model (such as XML-RoBERTa or mDeBERTa) was used or the data was automatically translated into the other language by using services such as Google Translate[8]. Among the participants, there were no approaches that went beyond these well-established procedures.

Since all participants pursued a supervised-learning approach, the choice of training data is also an issue that has been addressed by several participants. About half of them used some training data in addition to the one provided as part of the lab. A popular complement were data from *Kaggle*.[9] [10]

6.3. Detailed Description of Participants Systems

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

Team AI Rational (`monolingual:25 cross-lingual:8`) employed a RoBERTa classifier and made use of other English training data in addition to the provided dataset.

Team AIT_FHSTP [50] (`monolingual:22 cross-lingual:4`) primarily experimented with different transformers for this task being T5 and XLM-RoBERTa. For the evaluation, they used XLM-RoBERTa. For the cross-lingual subtask the given English training data was translated into German.

Team Awakened [41] (`monolingual:3 cross-lingual:5`) took part in both subtasks, employing a BiLSTM architecture with BART sentence transformers for the monolingual and BiLSTM with XLM sentence transformers for the cross-lingual task.

Team boby232 (`monolingual:11`) exclusively experimented with feature-based supervised classification, more specifically, k nearest neighbors. The focus was on tuning the parameters of the classifier. The training data provided by the previous edition of this task was used.

Team BUM [45] (`monolingual:10`) described an approach to the monolingual fake news detection task. Numerous additional datasets were added for training. In addition, a Bag-of-Words approach was used to extract text passages from Wikipedia data that match claims from the training and test data. A T5 transformer approach checked whether a claim was a logical consequence of these passages. As a result, the authors found that the approach worked better for detecting the *false* class than for the other classes. They attribute this to unbalanced data.

Team CIC [44] (`monolingual:8 cross-lingual:6`) considered three different classifiers: passive aggressive classification, Bi-LSTM and a transformer (i.e. RoBERTa). For the monolingual task, RoBERTa performed best, while for the cross-lingual task BiLSTM perform best.

[8] https://translate.google.com/

[9] www.kaggle.com/datasets/liberoliber/onion-notonion-datasets

[10] www.kaggle.com/c/fakenewskdd2020

Team CODE [47] (`monolingual:15`) built a system based on two components: the first component establishes whether an instance is relevant (i.e. not belonging to the *other* class) while the second component is devised to specify relevant instances (i.e. it distinguishes between the remaining class labels of this subtask). As component classifiers, the authors fine-tune BERT. The team employed additional training data: *Fake News Detection Challenge KDD 2020* and the *Fake News Classification Datasets from Kaggle.*

Team DIU_Carbine (`monolingual:14`) first augmented the dataset with additional instances of class *true* from Kaggle so that the dataset is more balanced for training. TF-IDF was used as a method to generate features for supervised learning. Four different traditional learning algorithms were tested, Logistic Regression turned out to be the one that worked best.

Team DIU_Fire71 (`monolingual:24`) experimented with several traditional learning algorithms, namely XGBoost, KNN, Gradient Boosting Classifier, Random Forest, Support Vector Machines, Naive Bayes, and Decision Trees. They also used other English training data than the one provided in this year's task and the dataset from the last iteration of the shared task. For the best classification model, they used TF-IDF vectorization. XGBoost and the Gradient Boosting Classifier algorithm achieved the best results.

Team DIU_Phoenix (`monolingual:21`) employed a multitude of traditional supervised learning algorithms (i.e. Support Vector Machines, Logistic Regression, Random Forests, Decision Trees, XGB Boosting, Gradient Boosting, Naive Bayes, KNN) and deep learning classfiers (LSTM). They used other English training data than the one provided in this year's task and the dataset from the last iteration of the shared task as well.

Team DIU_SilentKillers (`monolingual:23`) experimented with Support Vector Machines, Random Forests, XGBoost, and LSTM. No additional dataset was used.

Team DIU_SpeedOut (`monolingual:13`) experimented with several traditional learning algorithms, i.e. Naive Bayes, Logistic Regression, and Stochastic Gradient Descent, and deep learning, more specifically, LSTM.

Team FoSIL [48] (`monolingual:18`) employed an SVM in combination with a feature selection algorithm whose concept is based on human behaviour-based optimization.

Team HBDCI [46] (`monolingual:12`) compared two different classification approaches: a feature-based approach using traditional supervised learning and a deep-learning approach that combines BERT, CNN, non-contextual embeddings and stylometric features. Both classifiers were evaluated in different configurations (i.e. different subsets of features). Overall, the deep-learning approach outperformed the feature-based approach. Including a subset of stylometric features was also helpful.

Team iCompass (`monolingual:1`) employed two concatenated parallel fine-tuned BERT models. One of the models processed the title of an article and the other the main text.

Team MNB (`monolingual:16`) experimented with Support Vector Machines, Logistic Regression, Random Forests, and Naive Bayes. Tuning the parameters of their classifiers was the main focus of their research.

Team NITK-IT_NLP [40] (`monolingual:7 cross-lingual:2`) examined different transformer models for both subtasks. They also proposed a classifier trained on striding text windows of the data. This approach seems necessary since some of the document instances from the given dataset are fairly long.

Team NLP&IR@UNED [38] (monolingual:2) used first a longformer model to be able to process longer sequences as they are typical for news texts. As a second approach, they applied data augmentation by splitting the texts into shorter sequences before classifying with BERT-based models. As a third approach, the authors derived text and LIWC-features and used them together with a transformer embedding in an ensemble.

Team NLytics [42] (monolingual:5) experimented both with RoBERTa and Longformer models. They employed the latter for the official system submission in order to overcome the restriction of 512 tokens. A topic modeling approach was implemented prior to the classification step to account for the varying class distributions in different topics.

Team NoFake (cross-lingual:7) also made use of other English non-specified training data and training data provided by the previous edition of this task. They experimented with two traditional supervised classifiers (Support Vector Machines, Logistic Regression) and one BERT deep learning classifier. They focused on exploring the different training data in their research.

Team SCUoL [43] (monolingual:6) tested four supervised learning algorithms (features: TF-IDF) and four transformers. Additional data from the Kaggle task was experimented with. The result showed that SVC is the best classifier and bert-large-cased is the best transformer model, slightly outperforming SVC. However, the additional data did not result in any performance gains.

Team subMNB (monolingual:17) employed Support Vector Machines, Logistic Regression, Random Forests and Naive Bayes. In addition to the training data provided by in the context of the task, other English training data was used.

Team TextMinor [49] (monolingual:19) pursued a deep-learning approach based on RoBERT. Next to exploiting the information contained in that language model, the authors also included overlap features, a singular value decomposition between text and title, and cosine similarity between text and title based on their TF-IDF representation.

Team UNED (monolingual:4 cross-lingual:3) experimented with BERT, RoBERTa, and ALBERT deep learning classifiers. They relied on the following publicly available models: bert-base-cased, bert-base-uncased, albert-base-v2, multilingual-bert-base, roberta-base. Their focus was on the fine-tuning of those classifiers.

Team ur-iw-hnt [39] (monolingual:9 cross-lingual:1) experimented with extractive and abstractive summarization. Subsequently, BERT models were applied. In the case of German, the data was first automatically translated into English using machine translation. Large language models worked well, but overfitting was identified as an issue that needs be avoided.

7. Conclusion

We have presented a detailed overview of Task 3 of the CheckThat! Lab of CLEF 2022. It focused on the classification of news articles with respect to the correctness of their main claims. The results give a realistic estimate of the current state-of-the art for fake news detection. Most of the participants used transformer-based models like BERT or RoBERTa. Systems based on such technology could be applied within the fact checking community. However, the results show that more work is required in order to improve the current systems. The marco F1 scores are not sufficient for a satisfying multi-class classification of news articles according to their factuality. It is a limitation that the provided dataset was unbalanced. Yet, this shared task is one of the few research initiatives that focuses not binary, but multi-class classification. On top of that we offer a dataset in two languages: German and English. Future research should continue our efforts to provide high-quality multilingual real-world dataset in multiple languages and also broaden the scope by including different kinds of meta data (e.g. social factors). Thus, enabling research beyond textual features.

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.

Part of this work has been funded by the BMBF (German Federal Ministry of Education and Research) under the grant no. 01FP20031J. The responsibility for the contents of this publication lies with the authors.

References

[1] Y. Liu, Y.-F. B. Wu, Fned: A deep network for fake news early detection on social media, ACM Transactions on Information Systems 38 (2020) 1–33. doi:10.1145/3386253.

[2] Z. Wang, Z. Yin, Y. A. Argyris, Detecting medical misinformation on social media using multimodal deep learning, IEEE journal of biomedical and health informatics 25 (2021) 2193–2203. doi:10.1109/JBHI.2020.3037027.

[3] V. K. Singh, I. Ghosh, D. Sonagara, Detecting fake news stories via multimodal analysis, Journal of the Association for Information Science and Technology 72 (2021) 3–17. doi:10.1002/asi.24359.

[4] M. Villagracia Octaviano, Fake news detection using machine learning, in: 2021 5th International Conference on E-Society, E-Education and E-Technology, ICSET 2021, Association for Computing Machinery, New York, NY, USA, 2021, p. 177–180. URL: https://doi.org/10.1145/3485768.3485774. doi:10.1145/3485768.3485774.

[5] C.-C. Hsu, Y.-X. Zhuang, C.-Y. Lee, Deep fake image detection based on pairwise learning, Applied Sciences 10 (2020) 370. doi:10.3390/app10010370.

[6] S. Agarwal, H. Farid, T. El-Gaaly, S.-N. Lim, Detecting deep-fake videos from appearance and behavior, in: 2020 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2020, pp. 1–6. doi:10.1109/WIFS49906.2020.9360904.

[7] D. Kopev, A. Ali, I. Koychev, P. Nakov, Detecting deception in political debates using acoustic and textual features, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2019, pp. 652–659. doi:10.1109/ASRU46091.2019.9003892.

[8] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, S. Satoh, Spotfake: A multimodal framework for fake news detection, in: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), IEEE, 2019, pp. 39–47. doi:10.1109/BigMM.2019.00-44.

[9] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. S. Kartal, J. Beltrán, Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets, in: Working

Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[10] P. Nakov, G. Da San Martino, F. Alam, S. Shaar, H. Mubarak, N. Babulkov, Overview of the CLEF-2022 CheckThat! lab task 2 on detecting previously fact-checked claims, in: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[11] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, M. Wiegand, M. Siegel, J. Köhler, Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection, in: Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, CLEF '2022, Bologna, Italy, 2022.

[12] 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.

[13] 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.

[14] M. Hardalov, A. Arora, P. Nakov, I. Augenstein, A Survey on Stance Detection for Mis- and Disinformation Identification, in: Findings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL '22 (Findings), Association for Computational Linguistics, Seattle, WA, USA, 2022. URL: https://arxiv.org/abs/2103.00242.

[15] 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.

[16] A. Kazemi, K. Garimella, G. K. Shahi, D. Gaffney, S. A. Hale, Research note: Tiplines to uncover misinformation on encrypted platforms: A case study of the 2019 indian general election on whatsapp, Harvard Kennedy School Misinformation Review (2022).

[17] G. K. Shahi, J. M. Struß, T. Mandl, Overview of the CLEF-2021 checkthat! lab: Task 3 on fake news detection, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 406–423. URL: http://ceur-ws.org/Vol-2936/paper-30.pdf.

[18] M. Hardalov, I. Koychev, P. Nakov, In search of credible news, in: C. Dichev, G. Agre (Eds.), Artificial Intelligence: Methodology, Systems, and Applications, Springer International Publishing, Cham, 2016, pp. 172–180.

[19] G. K. Shahi, D. Nandini, Fakecovid–a multilingual cross-domain fact check news dataset for covid-19, arXiv preprint arXiv:2006.11343 (2020).

[20] D. Röchert, G. K. Shahi, G. Neubaum, B. Ross, S. Stieglitz, The networked context of covid-19 misinformation: Informational homogeneity on youtube at the beginning of the pandemic, Online Social Networks and Media 26 (2021) 100164.

[21] K. Pogorelov, D. T. Schroeder, S. Brenner, J. Langguth, Fakenews: Corona virus and 5g conspiracies multimedia analysis task at mediaeval 2021, in: Working Notes Proceedings of the MediaEval 2021 Workshop, 2022.

[22] 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: S. Hicks, D. Jha, K. Pogorelov, A. G. S. de Herrera, D. Bogdanov, P. Martin, S. Andreadis, M. Dao, Z. Liu, J. V. Quiros, B. Kille, M. A. Larson (Eds.), Working Notes Proceedings of the MediaEval 2020 Workshop, Online, 14-15 December 2020, volume 2882 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2882/paper64.pdf.

[23] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, A. Zubiaga, SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 69–76. URL: https://aclanthology.org/S17-2006. doi:10.18653/v1/S17-2006.

[24] 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, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 845–854. URL: https://aclanthology.org/S19-2147. doi:10.18653/v1/S19-2147.

[25] 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-2016), Association for Computational Linguistics, San Diego, California, 2016, pp. 31–41. URL: https://aclanthology.org/S16-1003. doi:10.18653/v1/S16-1003.

[26] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, P. Nakov, SemEval-2020 task 11: Detection of propaganda techniques in news articles, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1377–1414. URL: https://aclanthology.org/2020.semeval-1.186. doi:10.18653/v1/2020.semeval-1.186.

[27] 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, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 860–869. URL: https://aclanthology.org/S19-2149. doi:10.18653/v1/S19-2149.

[28] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fact extraction and VERification (FEVER) shared task, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1–9. URL: https://aclanthology.org/W18-5501. doi:10.18653/v1/W18-5501.

[29] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://aclanthology.org/N18-1074. doi:10.

18653/v1/N18-1074.

[30] W. Antoun, F. Baly, R. Achour, A. Hussein, H. Hajj, State of the art models for fake news detection tasks, in: 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), IEEE, 2020, pp. 519–524. doi:10.1109/ICIoT48696.2020.9089487.

[31] M. Mayank, S. Sharma, R. Sharma, DEAP-FAKED: knowledge graph based approach for fake news detection, CoRR abs/2107.10648 (2021). URL: https://arxiv.org/abs/2107.10648. arXiv:2107.10648.

[32] Q. Su, M. Wan, X. Liu, C.-R. Huang, et al., Motivations, methods and metrics of misinformation detection: an nlp perspective, Natural Language Processing Research 1 (2020) 1–13.

[33] 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.

[34] G. K. Shahi, T. A. Majchrzak, Exploring the spread of covid-19 misinformation on twitter, EasyChair Preprint no. 6009, EasyChair, 2021.

[35] G. K. Shahi, AMUSED: An annotation framework of multi-modal social media data, 2020. arXiv:2010.00502.

[36] G. K. Shahi, J. M. Struß, T. Mandl, J. Köhler, M. Wiegand, M. Siegel, CT-FAN-22 corpus: A Multilingual dataset for Fake News Detection, 2022. URL: https://doi.org/10.5281/zenodo.6555293. doi:10.5281/zenodo.6555293.

[37] B. Taboubi, M. A. B. Nessir, H. Haddad, iCompass at CheckThat! 2022: combining deep language models for fake news detection, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[38] J. R. Martinez-Rico, J. Martinez-Romo, L. Araujo, NLP&IRUNED at CheckThat! 2022: ensemble of classifiers for fake news detection, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[39] H. N. Tran, U. Kruschwitz, ur-iw-hnt at CheckThat! 2022: cross-lingual text summarization for fake news detection, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[40] R. L. Hariharan, M. Anand Kumar, Nitk-it_nlp at checkthat! 2022: Window based approach for fake news detection using transformers, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[41] C.-O. Truică, E.-S. Apostol, A. Paschke, Awakened at CheckThat! 2022: fake news detection using BiLSTM and sentence transformer, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[42] A. Pritzkau, O. Blanc, M. Geierhos, U. Schade, NLytics at CheckThat! 2022: hierarchical multi-class fake news detection of news articles exploiting the topic structure, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[43] S. Althabiti, M. A. Alsalka, E. Atwell, SCUoL at CheckThat! 2022: fake news detection using transformer-based models, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[44] M. Arif, A. L. Tonja, I. Ameer, O. Kolesnikova, A. Gelbukh, G. Sidorov, A. G. Meque, CIC

at CheckThat! 2022: multi-class and cross-lingual fake news detection, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[45] D. La Barbera, K. Roitero, J. Mackenzie, S. Damiano, G. Demartini, S. Mizzaro, BUM at CheckThat! 2022: a composite deep learning approach to fake news detection using evidence retrieval, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[46] C. P. Capetillo, D. Lecuona-Gómez, H. Gómez-Adorn, I. Arroyo-Fernández, J. Neri-Chávez, HBDCI at CheckThat! 2022: fake news detection using a combination of stylometric features and deep learning, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[47] O. Blanc, A. Pritzkau, U. Schade, M. Geierhos, CODE at CheckThat! 2022: multi-class fake news detection of news articles with BERT, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[48] A. Ludwig, J. Felser, J. Xi, D. Labudde, M. Spranger, FoSIL at CheckThat! 2022: using human behaviour-based optimization for text classification, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[49] S. Kumar, G. Kumar, S. R. Singh, TextMinor at CheckThat! 2022: fake news article detection using RoBERT, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[50] M. Schütz, J. Böck, M. Andresel, A. Kirchknopf, D. Liakhovets, D. Slijepčević, A. Schindler, AIT_FHSTP at CheckThat! 2022: Cross-lingual fake news detection with a large pre-trained transformer, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[51] M. Schütz, Detection and identification of fake news: Binary content classification with pre-trained language models, in: Information between Data and Knowledge, volume 74 of *Schriften zur Informationswissenschaft*, Werner Hülsbusch, Glückstadt, 2021, pp. 422–431. URL: https://epub.uni-regensburg.de/44959/, gerhard Lustig Award Papers.

[52] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn, The development and psychometricproperties of LIWC2015, Technical Report, University of Texas, 2015.