Overview of the CLEF-2021 CheckThat! Lab Task 2 on Detecting Previously Fact-Checked Claims in Tweets and Political Debates

Shaden Shaar[1], Fatima Haouari[2], Watheq Mansour[2], Maram Hasanain[2], Nikolay Babulkov[3], Firoj Alam[1], Giovanni Da San Martino[4], Tamer Elsayed[2] and Preslav Nakov[1]

[1]*Qatar Computing Research Institute, HBKU, Doha, Qatar*
[2]*Qatar University, Qatar*
[3]*Sofia University, Bulgaria*
[4]*University of Padova, Italy*

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 the *task 2*, which asks to detect previously fact-checked claims (in two languages). A total of four teams participated in this task, submitted a total of sixteen runs, and most submissions managed to achieve sizable improvements over the baselines using transformer based models such as BERT, RoBERTa. 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 datasets from the lab as well as the evaluation scripts, which should enable further research in detecting previously fact-checked claims.

Keywords

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

1. Introduction

There has been a growing concern about the spread of dis/mis-information in social media, ad this has become an urgent social and political issue. Over time, several initiatives for manual fact-checking have been launched, with over 200 fact-checking organizations actively working worldwide.[1] Unfortunately, these efforts do not scale and they are clearly insufficient, given the scale of disinformation propagating in different communication channels, which, according to the World Health Organization, has grown into the First Global Infodemic in the times of COVID-19.

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
sshaar@hbku.edu.qa (S. Shaar); 200159617@qu.edu.qa (F. Haouari); 200159617@qu.edu.qa (W. Mansour); maram.hasanain@qu.edu.qa (M. Hasanain); nbabulkov@gmail.com (N. Babulkov); fialam@hbku.edu.qa (F. Alam); dasan@math.unipd.it (G. D. S. Martino); telsayed@qu.edu.qa (T. Elsayed); pnakov@hbku.edu.qa (P. Nakov)


[1]http://tiny.cc/zd1fnz

There has been a surge in research to develop systems for automatic fact-checking. However, such systems suffer from credibility issues. Hence, it is important to reduce the manual effort by detecting when a claim has already been fact-checked. Work in this direction includes [1] and [2]: the former developed a dataset for the task and proposed a ranking model, while the latter proposed a neural ranking model using textual and visual modalities.

To deal with this problem, we launched the CheckThat! Lab, which features a number of tasks aiming to help automate the fact-checking process and to reduce the spread of disinformation and misinformation. The `CheckThat!` lab[2] was run for the fourth time in the framework of CLEF 2021. The purpose of the 2021 edition of the lab was to foster the development of technology that would enable finding check-worthy claims, finding claims that have been previously fact-checked, and predicting the veracity of a news article and its topic. Thus, the lab focuses on three types of content: (*i*) *tweets*, (*ii*) *political debates and speeches*, and (*iii*) *news articles*.

In this paper, we describe in detail the second task, *detecting previously fact-checked claims*, of the `CheckThat!` lab tasks.[3] Figure 1 shows the full `CheckThat!` identification and verification pipeline, including the tasks on detecting check-worthy claims, detecting previously fact-checked claims, and veracity and topic detection of news articles. The second task is defined as follows: *"given a check-worthy input claim and a set of verified claims, rank the previously verified claims in order of usefulness to fact-check the input claim."* It consists of the following two subtasks:

Subtask 2A: Detecting previously fact-checked claims in tweets. Given a tweet, detect whether the claim it makes was previously fact-checked with respect to a collection of fact-checked claims. This is a ranking task, offered in Arabic and English, where the systems need to return a list of top-n candidates.

Subtask 2B: Detecting previously fact-checked claims in political debates or speeches. Given a claim in a political debate or a speech, detect whether the claim has been previously fact-checked with respect to a collection of previously fact-checked claims. This is a ranking task, and it was offered in English.

For Subtask 2A, we focused on tweets, and it was offered in Arabic, and English. The participants were free to work on any language(s) of their interest, and they could also use multilingual approaches that make use of all datasets for training. Subtask 2A attracted four teams, and the most successful approaches used transformers or a combination of embeddings, manually engineered features, and neural networks. Section 3 offers more details.

For Subtask 2B, we focused on political debates and speeches, and we used PolitiFact as the main data source. The task attracted three teams, and a combination of transformers, prepossessing, and augmentation approaches performed the best. Section 4 gives more details.

As for the rest of the paper, Section 2 discusses some related work, and Section 5 concludes with final remarks.

[2] http://sites.google.com/view/clef2021-checkthat/

[3] Refer to [3] for an overview of the full CheckThat! 2021 lab.

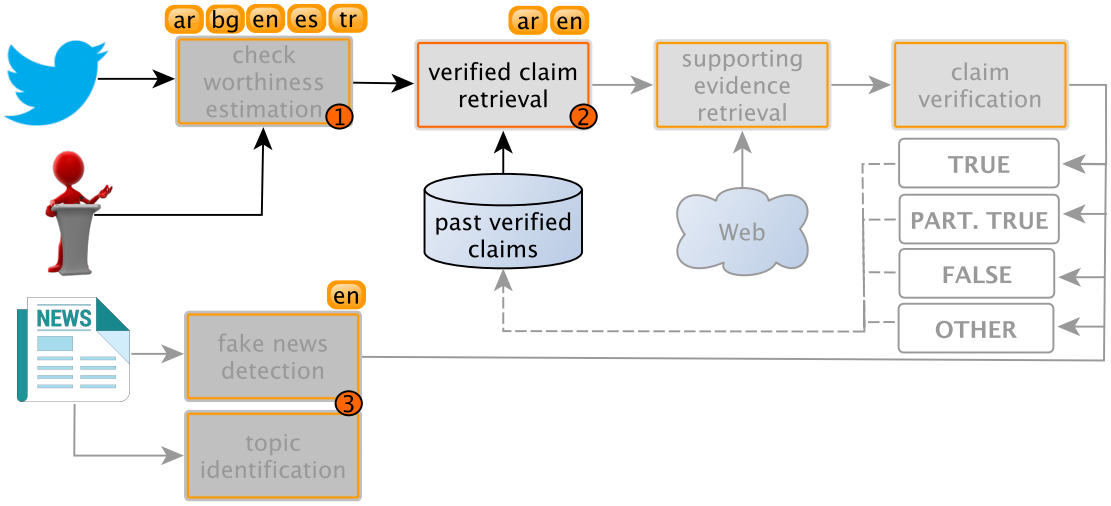

Figure 1: The full verification pipeline of the `CheckThat!` lab 2021. The verified claim retrieval subtask 2A targets tweets, and subtask 2B targets political debates. See [4, 5] for a discussion on tasks 1 and 3. The grayed tasks were addressed in previous editions of the lab [6, 7]

2. Related Work

A large body of research focused on developing automatic systems for fact-checking [8, 9, 10, 11, 12]. This includes datasets [13, 14], and evaluation campaigns [15, 6, 16, 17, 18]. However, there are credibility issues with automated systems [19], and thus a reasonable solution is to build tools to facilitate human fact-checkers, e.g., by detecting previously fact-checked claims.

This is an underexplored task and the only directly relevant work is [1, 20]; here, we use their annotation setup and one of their datasets: PolitiFact. Previous work has mentioned the task as an integral step of an end-to-end automated fact-checking pipeline, but there was very little detail provided about this component and it was not evaluated [21].

In an industrial setting, Google has developed the *Fact Check Explorer*,[4] which allows users to search a number of fact-checking websites. However, the tool cannot handle a complex claim, as it uses the standard Google search funcionality, which is not optimized for semantic matching of long claims.

Another related work is the *ClaimsKG* dataset and system [22], which includes 28K claims from multiple sources, organized into a knowledge graph (KG). The system can perform data exploration, e.g., it can find all claims that contain a certain named entity or keyphrase. In contrast, we are interested in detecting whether a claim was previously fact-checked.

Finally, the task is related to semantic relatedness tasks, e.g., from the GLUE benchmark [23], such as natural language inference (NLI) [24], recognizing textual entailment (RTE) [25], paraphrase detection [26], and semantic textual similarity (STS-B) [27]. However, it differs from them in a number of aspects; see [1] for more detail and discussion.

[4] http://toolbox.google.com/factcheck/explorer

3. Subtask 2A: Detecting Previously Fact-Checked Claims in Tweets

Given a tweet, the task asks to detect whether the claim the tweet makes was previously fact-checked with respect to a collection of fact-checked claims. The task is offered in Arabic and English. This is a ranking task, where the systems are asked to return a list of top-n candidates.

3.1. Dataset

Arabic To construct our verified claims collection, we selected 5,921 Arabic claims from AraFacts [28], and 24,408 English claims from ClaimsKG [29], which we translated to Arabic using the Google translate API.[5] To obtain our tweet–VerClaim pairs, we first selected a set of 1,274 Arabic verified claims from AraFacts such that each claim has at least one stated tweet example in its corresponding fact-checking article. Second, we selected one tweet example for each verified claim following the guidelines below:

1. Select an Arabic tweet.
2. Avoid tweets where the claim is stated in an image or a video.
3. Try to choose the tweet example that does not exactly match the text of the claim.
4. Avoid tweets that are relevant, but do not contain the claim or it is not clear whether they are about the claim.
5. Avoid tweets that have more than one claim.

The two annotators who constructed the tweet–VerClaim pairs swapped their pairs to double-check that the selected tweets were compliant with the guidelines. They further resolved any disagreements by discussing the reasons behind their choice, and excluded the claims where the disagreement remains. We ended up with 858 tweet–VerClaim pairs.

Due to the fact that AraFacts contains verified claims from five different Arabic fact-checking platforms, and since a claim can be verified by multiple sources, we had to check whether the annotated tweets can be paired with more than one claim from our verified claims collection. We first adopted Jaccard similarity to check whether each verified claim in our collection was verified by multiple sources. For each given verified claim, we selected all the claims that had a Jaccard similarity above 30%; then, we asked the annotators to double-check and to exclude any non-similar claims. Given similar claims to the ones in our qrels (query relevances), we constructed new tweet–VerClaim pairs.

To further verify any missing similar claims, we used Pyserini [30] to index our verified claims collection and we retrieved the top-25 potentially relevant verified claims for each tweet in our dataset. One annotator then checked for missing tweet–VerClaim pairs in our previously constructed qrels, and we expanded the qrels accordingly. Figure 2 presents some input tweet examples from our dataset and the corresponding top-5 verified claims ranked based on their relevance using a BM25 system.

[5] https://cloud.google.com/translate

Figure 2: Task 2A, Arabic: Examples of input tweets and the top-5 most similar verified claims from our verified claims collection retrieved by a BM25 system. Correct matches to previously verified matching claims are marked with a ✓.

Input tweet	(a)	*سرقة بنك في أمريكا و توزيع الفلوس على المارة*
Verified claims	**(1)**	✓ **أمريكا: سرقة بنك وتوزيع النقود على الناس (فيديو)**
	(2)	✖ قرر الملك عبد الله الثاني بن الحسين بالتعاون مع بنك الأردن وبسبب جائحة كورونا، التي مرت على العالم العربي والعالم أجمع، توزيع منح مالية علي المواطنين
	(3)	✖ نصب خيم و توزيع الحلويات احتفالاً بخبر اصابة النائب جبران باسيل بفيروس كورونا
	(4)	✓ **متظاهرون في أميركا يستولون على بنك ويوزّعون النقود على الناس**
	(5)	✖ (فيديو) توزيع وزيرة الصحة في كوريا الجنوبية راتبها على الفقراء
Input tweet	(b)	*هنا #برلين تطالب بسحب اللاجئين من إدلب إلى ألمانيا هؤلاء يعلمون أن ما يحدث في سورية ليس حرباً أهلية ولا حرباً طائفية ولا ثورة جياع إنما هو ثورة شعبية وطنية حقوقية يعلمون أكثر من السوريين نفسهم المحسوبين على الثورة مليون ونصف سوري ببرلين ولا سوري بخيمه #عرسال_تستغيث*
Verified claims	(1)	✖ في إدلب ولا في القامشلي
	(2)	✓ **مظاهرة في ألمانيا تطالب باستقبال السوريين الذين يعيشون في إدلب تحت القصف لإنقاذهم وإعطائهم حق اللجوء.**
	(3)	✖ بإيران الثورة عندهم ما فيها مزح ثورة جحاشيية
	(4)	✖ الفاتيكان يعثر على نسخة قديمة من الإنجيل ويصدم العالم المسيحي بأنّ عيسى ليس الله ولا ابناً لله.
	(5)	✖ امرأة مسنة يمنية ماتت دون أن يسأل عنها أحد، ومرّ على موتها شهور. في رواية أخرى تم العثور على امرأة سورية من ادلب تعيش وحدها في اسطنبول ميتة في منزلها. وقال البعض أن المرأة ميتة في منزلها منذ أكثر من عام في مدينة إدلب السورية.

English To construct the verified claims database, we used Snopes, a fact-checking website that targets rumors spreading in social media, and we collected 13,835 verified claims. Their fact-checking journalists often cite the tweet or the social media post that spreads the rumor when writing an article about a claim. We looked over all crawled verified articles, and we collected 1,401 tweets.

Table 1 shows examples of the input tweets and the results retrieved by the BM25 baseline. From example 1 of the table, we can see that the tweet–vclaim pairs are more complex than simple textual similarity. Then, example 2 shows that, in order to do a good decision about a pair, we need to understand the contextual meaning of the sentences.

Table 2 shows statistics about the CT–VCR–21 corpus for Task 2, including both subtasks and languages. CT–VCR–21 stands for `CheckThat!` verified claim retrieval 2021. *Input–VerClaim* pairs represent input claims with their corresponding verified claims by a fact-checking source. For Arabic, we randomly split the data into 512 training, 85 development, and 261 test examples. In total, the Arabic dataset consists of 858 queries, 1,039 qrels, and a collection of 30,329 verified claims. For English, we split the data into 70%, 15% and 15% for training, development, and test, respectively.

3.2. Evaluation

For the ranking tasks, as in the two previous editions of the `CheckThat!` lab, we calculated *Mean Average Precision* (MAP), reciprocal rank, Precision@k ($P@k$) and MAP@k for $k \in \{1, 3, 5, 10, 20, 30\}$. We used MAP@5 as the official evaluation measure.

Table 1
Task 2A, English: Examples of input tweets and the top-5 most similar verified claims from our verified claims collection retrieved by a BM25 system. The correct previously verified matching claims to be retrieved are marked with a ✓.

Input tweet	*(a)*	*Sen. Mitch McConnell: "As recently as October, now-President Biden said you can't legislate by executive action unless you are a dictator. Well, in one week, he signed more than 30 unilateral actions." pic.twitter.com/PYQKe9Geez — Forbes (@Forbes) January 28, 2021*	
Verified claims	**(1)**	**When he was still a candidate for the presidency in October 2020, U.S. President Joe Biden said, "You can't legislate by executive order unless you're a dictator."**	✓
	(2)	Photographs you post on Snapchat can now be used as evidence in legal cases unless you opt out.	✗
	(3)	U.S. Sen. Mitch McConnell said he would not participate in 2020 election debates that include female moderators.	✗
	(4)	U.S. Sen. Majority Leader Mitch McConnell said that U.S. President Trump "provoked" the attack on the Capitol.	✗
	(5)	President Joe Biden signed an executive order in 2021 allowing the U.S. to fund abortions abroad.	✗
Input tweet	*(b)*	*A supporter of President Donald Trump carries a Confederate battle flag on the second floor of the U.S. Capitol near the entrance to the Senate after breaching security defenses, in Washington, January 6, 2021. Photo by Mike Theiler pic.twitter.com/pbhwfAVsUX — corinne_perkins (@corinne_perkins) January 6, 2021*	
Verified claims	(1)	In January 2021, Hillary Clinton suggested U.S. President Donald Trump spoke by phone with Vladimir Putin on the day of an attack on the U.S. Capitol, Jan. 6, 2021.	✗
	(2)	In January 2021, OnlyFans removed Donald Trump's account in the aftermath of the Jan. 6 attack on the U.S. Capitol.	✗
	(3)	**A Confederate flag was spotted inside and outside the U.S. Capitol as a pro-Trump mob stormed the building.**	✓
	(4)	A pro-Trump mob chanted "Hang Mike Pence" as they stormed the U.S. Capitol on Jan. 6, 2021.	✗
	(5)	Kevin Seefried, who carried a Confederate flag into the U.S. Capitol during the attack on the building in January 2021, is registered as a Democrat in Delaware.	✗

3.3. Overview of the Systems

A total of four teams participated in this task, submitting sixteen runs. One team participated in the Arabic task and three teams participated in the English task. Below, we discuss briefly the approach of each team.

Team bigIR (2A:ar:1) fine-tuned AraBERT [31] by adding two neural network layers on top of it to predict the relevance score for a given tweet–VerClaim pair. The fine-tuned model was used to re-rank the candidate claims based on the predicted relevance scores.

Table 2
Task 2: Statistics about the CT–VCR–21 corpus, including the number of *Input–VerClaim* pairs and the number of *VerClaim* claims to match a claim against.

	2A–Arabic	2A–English	2B–English
Input claims	**858**	**1,401**	**669**
Training	512	999	472
Development	85	200	119
Test	261	202	78
Input–*VerClaim* pairs	**1,039**	**1,401**	**804**
Training	602	999	562
Development	102	200	139
Test	335	202	103
Verified claims (to match against)	**30,329**	**13,835**	**19,250**

Team Aschern [32] (2A:en:1) used TF.IDF, fine-tuned pre-trained sentence-level BERT, and the re-ranking LambdaMART model. The system is evaluated on the English version of the dataset collected from tweets.

Team NLytics (2A:en:2) used RoBERTa with a regression function in the final layer by considering the problem as a ranking task.

Team DIPS [33] (2A:en:3) used Sentence-BERT embeddings for all claims and then computed the cosine similarity for each pair of an input tweet and a verified claim. The prediction was made by passing a sorted list of cosine similarities to a neural network.

3.4. Results

Table 3 shows the official evaluation results for subtask 2A for Arabic and for English. We can see that all four participating teams managed to outperform the corresponding Elastic Search (ES) baseline, which is actually a strong baseline.

Arabic A single system was submitted for this task by the bigIR team. They used AraBERT to re-rank a list of candidates retrieved by a BM25 model. They first constructed a balanced training dataset where the positive examples correspond to the query relevance (qrels) provided by the organizers, while the negative examples were selected from the top retrieved candidates by BM25 such that they are not already labeled as positive. Second, they fine-tuned AraBERT to predict the relevance score for a given tweet–VerClaim pair. They added two neural network layers on top of AraBERT to perform the classification. Finally, at inference time, they used BM25 to retrieve the top 20 candidate verified-claims. Then, they fed each tweet–VerClaim pair to the fine-tuned model to obtain a relevance score and to re-rank the candidate claims accordingly. Their system outperformed the Elastic Search baseline by a sizable margin achieving a MAP@5 of 0.908 (compared to 0.794 for Elastic Search baseline).

Table 3
Task 2A: Official evaluation results, in terms of MRR, MAP@k, and Precision@k. The teams are ranked by the official evaluation measure: MAP@5. Here, *ES baseline* is the Elastic Search baseline, which implements BM25.

	Team	MRR	MAP				Precision				
			@1	@3	@5	@10	@20	@1	@3	@5	@10	@20
Arabic												
1	bigIR	0.924	0.787	0.905	**0.908**	0.910	0.912	0.908	0.391	0.237	0.120	0.061
2	ES-baseline	0.835	0.682	0.782	**0.794**	0.799	0.802	0.793	0.344	0.217	0.113	0.058
English												
1	Aschern [32]	0.884	0.861	0.880	**0.883**	0.884	0.884	0.861	0.300	0.182	0.092	0.046
2	NLytics [34]	0.807	0.738	0.792	**0.799**	0.804	0.806	0.738	0.289	0.179	0.093	0.048
3	DIPS [33]	0.795	0.728	0.778	**0.787**	0.791	0.794	0.728	0.282	0.177	0.092	0.048
	ES baseline	0.761	0.703	0.741	**0.749**	0.757	0.759	0.703	0.262	0.164	0.088	0.046

English Three teams participated for English, submitting a total of ten runs. All of them managed to improve over the Elastic Search (ES) baseline by a large margin. Team **Aschern** performed best; they used TF.IDF, fine-tuned pre-trained sentence-BERT, and LambdaMART for re-ranking, and scored 13.4 (MAP@5) points above the baseline. The second-best system was submitted by the **NLytics** team, which fine-tuned RoBERTa, improving by 5 (MAP@5) points absolute over the baseline.

4. Subtask 2B: Detecting Previously Fact-Checked Claims in Political Debates or Speeches

Given a claim in a political debate or a speech, the task asks to detect whether the claim has been previously fact-checked with respect to a collection of previously fact-checked claims. This is also a ranking task, and it was offered in English.

4.1. Dataset

We have 669 claims from political debates [1], matched against 804 verified claims (some input claims match more than one verified claim) in a collection of 19,250 verified claims in PolitiFact. We report some statistics about the dataset in the last column of Table 2.

4.2. Evaluation

Similarly to subtask-2A, we treat this as a ranking task, and we report the same evaluation measures. Once again, MAP@5 is the official evaluation measure.

Table 4
Task 2b, English: Example of input claims and the top-5 most similar verified claims from our verified claims collection retrieved by a BM25 system. The correct previously verified matching claims to be retrieved are marked with a ✓.

Input tweet	*(a)*	*Richard Nixon released tax returns when he was under audit.*	
Verified claims	**(1)**	**Richard Nixon released tax returns when he was under audit.**	✓
	(2)	Every Republican nominee since Richard Nixon, who at one time was under an audit, has released their tax returns.	✗
	(3)	Even Richard Nixon released his tax returns to the public when he was running for president ...	✗
	(4)	Richard Nixon was the last president to be impeached.	✗
	(5)	Says his campaign has released his past tax returns.	✗
Input tweet	*(b)*	*He actually advocated for the actions we took in Libya and urged that Gadhafi be taken out, after actually doing some business with him one time.*	
Verified claims	(1)	If you actually took the number of Muslims [sic] Americans, we'd be one of the largest Muslim countries in the world.	✗
	(2)	Theres only one of us whos actually cut government spending not two, theres one and youre looking at him.	✗
	(3)	Says Roy Moore "has advocated getting the federal government out of health care altogether, which means doing away with Medicaid, which means doing away with Medicare."	✗
	(4)	**Says Donald Trump is "on record extensively supporting (the) intervention in Libya."**	✓
	(5)	**"When Moammar Gadhafi was set to visit the United Nations, and no one would let him stay in New York, Trump allowed Gadhafi to set up an elaborate tent at his Westchester County (New York) estate."**	✓

Table 5
Task 2B (English): Official evaluation results, in terms of MAP, MAP@k, and Precision@k. The teams are ranked by the official evaluation measure: MAP@5.

	Team	**MRR**	**MAP**					**Precision**			
			@1	**@3**	**@5**	**@10**	**@20**	**@1**	**@3**	**@5**	**@10**	**@20**
	ES-baseline	0.350	0.304	0.339	**0.346**	0.351	0.353	0.304	0.143	0.091	0.052	0.027
1	DIPS [33]	0.336	0.278	0.313	**0.328**	0.338	0.342	0.266	0.143	0.099	0.059	0.032
2	Beasku [35]	0.320	0.266	0.308	**0.327**	0.332	0.332	0.253	0.139	0.101	0.056	0.028
3	NLytics [34]	0.216	0.171	0.210	**0.215**	0.219	0.222	0.165	0.101	0.068	0.038	0.022

4.3. Overview of the Systems

Among the three participating teams, none could beat the official baseline. Below, we offer a short description of each systems.

Team DIPS [33] (2B:en:2) was the top-ranked team. They used sentence -BERT embeddings for all claims (input and verified), then computed a cosine similarity for each pair of an input claim and a verified claim. Finally, they made a prediction by passing a sorted list of cosine similarities to a neural network.

Team BeaSku [35] (2B:en:3) used triplet loss training to fine-tune sentence BERT. Then, they used the scores predicted by that model along with BM25 scores as features to train a rankSVM re-ranker. They further studied the impact of applying online mining of triplets, and they performed some experiments to augment the dataset automatically.

Team NLytics (2B:en:4) fine-tuned RoBERTa with a regression function in the final layer, treating the problem as a ranking task.

4.4. Results

Table 5 shows the official results for Task 2B. We can see that only three teams participated in this subtask, submitting a total of five runs, and no team managed to outperform the Elastic Search (ES) baseline, which is based on BM25.

5. Conclusion and Future Work

We have provided a detailed overview of the CLEF 2021 `CheckThat!` lab task 2, which focused on detecting previously fact-checked claims in tweets (Subtask 2A), and in political debates or speeches (Subtask 2B). Inline with the general mission of CLEF, we promoted multi-linguality by offering the task in two different languages: Arabic and English. The participating systems fine-tuned transformer models (such as BERT and RoBERTa) and some tried data augmentation. For Subtask 2A, four systems (one for Arabic and three for English) participated, and all outperformed a BM25 baseline. For Subtask 2B, none of the three participating teams could beat the baseline.

We plan a new iteration of the CLEF `CheckThat!` lab and of task 2, which will offer new larger training datasets and additional languages.

Acknowledgments

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

This work is part of the Tanbih mega-project,[6] developed at the Qatar Computing Research Institute, HBKU, which aims to limit the impact of "fake news", propaganda, and media bias by making users aware of what they are reading, thus promoting media literacy and critical thinking.

[6] http://tanbih.qcri.org

References

[1] S. Shaar, N. Babulkov, G. Da San Martino, P. Nakov, That is a known lie: Detecting previously fact-checked claims, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, 2020, pp. 3607–3618.

[2] N. Vo, K. Lee, Where are the facts? searching for fact-checked information to alleviate the spread of fake news, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, 2020, pp. 7717–7731.

[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, CLEF 2021, 2021.

[4] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, A. Nikolov, M. Kutlu, Y. S. Kartal, F. Alam, G. Da San Martino, A. Barrón-Cedeño, R. Míguez, T. Elsayed, P. Nakov, Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates, in: [36], 2021.

[5] G. K. Shahi, J. M. Struß, T. Mandl, Overview of the CLEF-2021 CheckThat! lab: Task 3 on fake news detection, in: [36], 2021.

[6] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, S. Shaar, Z. Sheikh Ali, Overview of CheckThat! 2020: Automatic identification and verification of claims in social media, LNCS (12260), 2020.

[7] T. Elsayed, P. Nakov, A. Barrón-Cedeño, M. Hasanain, R. Suwaileh, G. Da San Martino, P. Atanasova, Overview of the CLEF-2019 CheckThat!: Automatic identification and verification of claims, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, LNCS, 2019, pp. 301–321.

[8] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, J. Han, A survey on truth discovery, SIGKDD Explor. Newsl. 17 (2016) 1–16.

[9] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, SIGKDD 19 (2017) 22–36.

[10] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, J. L. Zittrain, The science of fake news, Science 359 (2018) 1094–1096.

[11] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151.

[12] N. Vo, K. Lee, The rise of guardians: Fact-checking URL recommendation to combat fake news, in: Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018, 2018, pp. 275–284.

[13] N. Hassan, C. Li, M. Tremayne, Detecting check-worthy factual claims in presidential debates, in: J. Bailey, A. Moffat, C. C. Aggarwal, M. de Rijke, R. Kumar, V. Murdock, T. K. Sellis, J. X. Yu (Eds.), Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015, 2015, pp. 1835–1838.

[14] I. Augenstein, C. Lioma, D. Wang, L. Chaves Lima, C. Hansen, C. Hansen, J. G. Simonsen, MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, 2019, pp. 4685–4697.
[15] J. Thorne, A. Vlachos, Automated fact checking: Task formulations, methods and future directions, in: COLING, 2018, pp. 3346–3359.
[16] S. Shaar, A. Nikolov, N. Babulkov, F. Alam, A. Barrón-Cedeño, T. Elsayed, M. Hasanain, R. Suwaileh, F. Haouari, G. Da San Martino, P. Nakov, Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media, in: [37], 2020.
[17] M. Hasanain, F. Haouari, R. Suwaileh, Z. Ali, B. Hamdan, T. Elsayed, A. Barrón-Cedeño, G. Da San Martino, P. Nakov, Overview of CheckThat! 2020 Arabic: Automatic identification and verification of claims in social media, in: [37], 2020.
[18] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. Kishore Shahi, J. Maria Struß, T. Mandl, The CLEF-2021 CheckThat! Lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: ECIR, 2021, pp. 639–649.
[19] P. Arnold, The challenges of online fact checking, Technical Report, Full Fact, 2020.
[20] S. Shaar, F. Alam, G. D. S. Martino, P. Nakov, The role of context in detecting previously fact-checked claims, arXiv:2104.07423 (2021).
[21] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, V. Sable, C. Li, M. Tremayne, ClaimBuster: The first-ever end-to-end fact-checking system, Proceedings of VLDB Endow. 10 (2017) 1945–1948.
[22] 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: Proceedings of the 18th International Semantic Web Conference, ISWC 2019, 2019, pp. 309–324.
[23] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, 2019.
[24] A. Williams, N. Nangia, S. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, 2018, pp. 1112–1122.
[25] L. Bentivogli, I. Dagan, H. T. Dang, D. Giampiccolo, B. Magnini, The fifth PASCAL recognizing textual entailment challenge, in: Proceedings of the Text Analysis Conference, TAC '09, 2009.
[26] W. B. Dolan, C. Brockett, Automatically constructing a corpus of sentential paraphrases, in: Proceedings of the Third International Workshop on Paraphrasing, 2005.
[27] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, L. Specia, SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation, in: Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval 2017, 2017, pp. 1–14.
[28] Z. S. Ali, W. Mansour, T. Elsayed, A. Al-Ali, AraFacts: The first large Arabic dataset of naturally occurring claims, in: Proceedings of the Sixth Arabic Natural Language Processing Workshop, 2021, pp. 231–236.

[29] 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: Proceedings of the International Semantic Web Conference, ISWC 2019, Springer, 2019, pp. 309–324.
[30] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, R. Nogueira, Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations, in: Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2021.
[31] W. Antoun, F. Baly, H. Hajj, AraBERT: Transformer-based model for Arabic language understanding, in: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, OSAC '20, Marseille, France, 2020, pp. 9–15.
[32] A. Chernyavskiy, D. Ilvovsky, P. Nakov, Aschern at CLEF CheckThat! 2021: Lambda-calculus of fact-checked claims, in: [36], 2021.
[33] S. Mihaylova, I. Borisova, D. Chemishanov, P. Hadzhitsanev, M. Hardalov, P. Nakov, DIPS at CheckThat! 2021: Verified claim retrieval, in: [36], 2021.
[34] A. Pritzkau, NLytics at CheckThat! 2021: Multi-class fake news detection of news articles and domain identification with RoBERTa - a baseline model, in: [36], 2021.
[35] B. Skuczyńska, S. Shaar, J. Spenader, P. Nakov, BeaSku at CheckThat! 2021: Fine–Tuning Sentence BERT with Triplet Loss and Limited Data, in: [36], 2021.
[36] G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Working Notes. Working Notes of CLEF 2021–Conference and Labs of the Evaluation Forum, 2021.
[37] L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), CLEF 2020 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2020.