Fraunhofer SIT at CheckThat! 2022: Ensemble Similarity Estimation for Finding Previously Fact-Checked Claims

Raphael Antonius Frick[1], Inna Vogel[1]

[1]*Fraunhofer Institute for Secure Information Technology SIT | ATHENE - National Research Center for Applied Cybersecurity*
Rheinstrasse 75, Darmstadt, 64295, Germany
https://www.sit.fraunhofer.de/

Abstract

During the corona pandemic misinformation has been increasingly spread on social media. Since the automatic verification of social media postings has shown to be challenging, there exists the need of systems, that can identify whether a claim in a post has already been previously analyzed by independent fact-checkers. In this paper, a system based on ensemble classification is proposed. It takes advantage of state-of-the-art sentence transformers for estimating the semantic similarity between a given tweet and individual parts of a fact-check. Furthermore, it incorporates several preprocessing steps as well as back-translation as a data augmentation technique. The proposed model ranked sixth best in the competition.

Keywords

Similarity Estimation, Sentence Transformer, Ensemble Classification, Twitter

1. Introduction

During the corona pandemic, a lot of misinformation was distributed over social media and instant messengers. They were used to sharing conspiracy theories about the existence of the virus, the origin of the outbreak as well as to promoting alternative medication in favor of the vaccine. Thus, there exists the need to detect, whether a social media posting contains false information or claims.

Since automatic verification of social media content still cannot be done with high accuracy [1], they aren't reliable enough to be used in practice e.g., by journalists. They are required to conduct fact-checking prior to publishing any article. Journalists sometimes leverage from information collected from social media due to them being the only source available. Major media houses, such as the AFP, Reuters and DPA, either have an in-house department or rely on third-party NGOs, such as Snopes, checking statements manually. Some of these fact-checks are made publicly available and can help with estimating the truthfulness of a claim.

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

raphael.frick@sit.fraunhofer.de (R. A. Frick); inna.vogel@sit.fraunhofer.de (I. Vogel)

The subtask 2A of this year's CLEF2022-CheckThat! Challenge [2, 3] revolved around detecting, whether a statement made in a tweet was previously fact-checked. In this paper, a new approach to identifying whether the content of a tweet has already been subject to fact-checking is presented. It combines multiple state-of-the-art sentence transformers for estimating the semantic similarity between tweets and fact-checks. The similarity score is then used to rank the fact-checks based on the likelihood, that they cover the statements and claims made in a given tweet.

The paper remainder is structured as follows: at first, an introduction to related work presented in the previous iteration of this competition is given in Section 2. Then, in Section 3 the task and the associated datasets are presented. In Section 4 an overview of the proposed approach is given along with explanations on the applied data augmentation and preprocessing. Section 5 showcases and discusses the experimental results achieved on the test set. The paper then concludes in Section 6 with an outlook for future work.

2. Related Work

The identification of previously fact-checked claims was part of the previous iteration of the CheckThat! Challenge [4, 5]. The best performing system utilized a combination of sentence-BERT and TF-IDF as features [6]. Furthermore, the author took advantage of LambdaMART for re-ranking the outputs. The second best submission [7] fine-tuned a RoBERTa model to be used to solve the ranking as a regression problem, while the system that ranked third [8], also utilized a sentence-BERT, but used a neural network for re-ranking.

3. Task & Dataset Description

The objective of this year's CheckThat! Lab task 2 [3] was to identify whether a claim had already been subject to a previously conducted fact-check. The task itself was divided into several subtasks, with the objective of task 2A being to predict, which previously fact-checked claim is represented by a given tweet. Although the task was available in English and in Arabic, the authors only participated in the English variant of the task.

Along with the challenge a dataset containing $14,231$ fact checks collected from Snopes and a multiset of tweets had been released. The tweets were split into a train split, a validation split, a validation-test split and a test split. Statistics on the number of tweets featured in each split is showcased in Table 2. Each fact-check consisted of three main parts: a title, a subtitle and the claim as visualized in Table 1. Further information contained the date of publishing as well as the author. While each tweet only referenced exactly one fact-check, one fact-check may refer to one or more tweets.

4. Methodology

In this section, the proposed ranking pipeline is presented. The proposed concept is visualized in Figure 1.

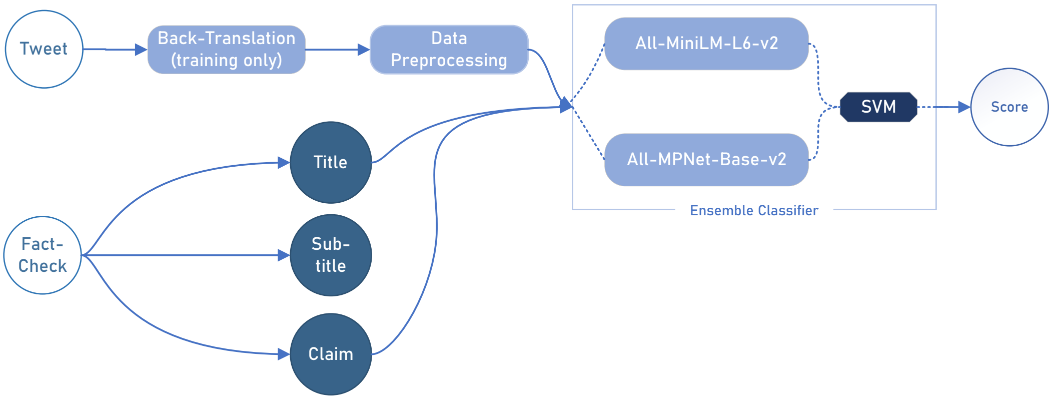
Table 1
Example of the structure of a fact-check conducted by *Snopes* in the dataset

Title	Did Vandals Attack 12 Churches in France in One Week?
Subtitle	A Breitbart report re-emerged on social media in the aftermath of the April 2019 fire at Notre Dame Cathedral in Paris.
Claim	In early 2019, vandals targeted 12 churches in France over seven days.

Table 2
Number of tweets per split of the CheckThat! 2022 Task 2A English dataset

	#Tweets
Train Split	999
Validation Split	200
Validation-Test Split	202
Test Split	209

Figure 1: Visualization of the ranking pipeline



4.1. Data Augmentation

Back-translation was used to augment training input data so that the models learn to generalize better to unseen data and to generate additional positive samples; samples which are semantically identical to a previously conducted fact-check. Back-translation allows creating new data samples from a reference sample by translating the content of the reference sample into a target language and then back into the language of the reference. The newly crafted samples feature a similar semantic to reference sample, but will differ a lot syntactically. For this lab, back-translation using German, Spanish and Chinese was considered. Translating sentences into Chinese may result in more translation errors than e.g., a translation into German. This however, may allow simulating the introduction of typos and grammar errors that are often found in social media posts. Google Translate hereby served as the translation service and the resulting samples were then used to

Table 3
Example of the structure of a fact-check conducted by *Snopes* in the dataset

Original	A number of fraudulent text messages informing individuals they have been selected for a military draft have circulated throughout the country this week.
German	A number of fraudulent text messages that have been selected throughout the country for a military design have been selected throughout the country.
Spanish	Several fraudulent text messages that report people who have been selected for a military draft have circulated throughout the country this week.
Chinese	Many fraudulent SMS notifications that they have been selected as a military draft individuals have been scattered throughout the country this week.

Table 4
Result of the preprocessing applied on a sample tweet

Original	This is the content we need #coronavirus content 😂 😂 🇺🇸 Https://t.co/gwjzukb16u
Preprocessed	This is the content we need coronavirus content emoji face with tears of joy emoji emoji face with tears of joy emoji emoji United States emoji

enrich the training dataset. Example results are showcased in Table 3.

4.2. Preprocessing

The tweets in the dataset need to undergo several preprocessing steps, before can be used in the ranking system. Preprocessing is mainly done with the help of the Python package pysentimiento[9]. Firstly, emojis are converted into descriptive tokens. Secondly, user mentions (e.g., @reuters) and URLs are turned into generic tokens. The generic URL tokens are then removed as part of the preprocessing step, as they do not contain any useful semantic information. The result on a sample is displayed in Table 4.

4.3. Ensemble Classifier

The ensemble classifier consists of weak classifiers, which outputs were merged to a unified similarity score using a meta-classifier. The base for the weak classifiers consisted of fine-tuned all-MiniLM-L6-v2 and all-MPNet-Base-v2 sentence transformers [10]. While experiments had also considered the utilization of other sentence-transformer models, all-MiniLM-L6-v2 and all-MPNet-Base-v2 scored best on the validation set. For each part of the fact-check (title, subtitle, claim) a classifier was trained for 10 epochs on the augmented train set. Hereby, the adam-optimizer was used in conjunction with an inital learning rate of 0.00002. The training set consisted of a randomly selected subset of the released train split. For each positive sample, a tweet and its corresponding fact-check, a negative pair using the same tweet and a fact-check not covering the claim has been crafted. This resulted in a balanced dataset and by utilizing the back-translated samples during pair generation, more negative samples could be considered during training. The training procedure took advantage of model checkpoints. By this only the best performing models on the validation set were kept, resulting in a total of $2 \times 3 = 6$ weak classifiers.

Table 5
Scored Reciprocal Rank, Precision and MAP of the individual models on the test set.
* indicates the submitted model

		All-MiniLM-L6-v2 (Claims)	All-MPNet-Base-v2 (Claims)	All-MiniLM-L6-v2 (Title)	All-MPNet-Base-v2 (Title)	All-MiniLM-L6-v2 (Subtitle)	All-MPNet-Base-v2 (Subtitle)
Reciprocal Rank		0.7614	0.7105	0.7380	0.7028	0.2614	0.2995
Prec. @N	@1	0.7143	0.6476	0.6714	0.6381	0.1905	0.2476
	@3	0.2635	0.2492	0.2635	0.2460	0.0968	0.1000
	@5	0.1629	0.1571	0.1619	0.1533	0.0667	0.0686
	@10	0.0833	0.0833	0.0829	0.0819	0.0395	0.0410
	@20	0.0436	0.0426	0.0426	0.0417	0.0231	0.0245
Map @N	@1	0.7143	0.6476	0.6714	0.6381	0.1905	0.2476
	@3	0.7484	0.6921	0.7286	0.6857	0.2349	0.2722
	@5	0.7534	0.7009	0.7326	0.6921	0.2452	0.2817
	@10	0.7565	0.7069	0.7352	0.7000	0.2535	0.2907
	@20	0.7595	0.7082	0.7368	0.7009	0.2581	0.2960

		Similarity Score Averaging	Meta Classifier	Meta Classifier No Preprocessing	Meta Classifier No Augmentation
		(Claim + Title + Subtitle)*	(Claim + Title)	(Claim + Title)	(Claim + Title)
Reciprocal Rank		0.6236	0.8014	0.7980	0.7979
Prec. @N	@1	0.5571	0.0.7524	0.7476	0.7614
	@3	0.2206	0.2746	0.2778	0.2678
	@5	0.1410	0.1705	0.1705	0.1614
	@10	0.0752	0.0881	0.0876	0.0861
	@20	0.0407	0.0450	0.0450	0.0442
Map @N	@1	0.5571	0.7524	0.7476	0.7614
	@3	0.6008	0.7873	0.7873	0.7868
	@5	0.6103	0.7942	0.7916	0.7911
	@10	0.6167	0.7983	0.7945	0.7940
	@20	0.6207	0.7999	0.7963	0.7976

Since sentence transformers output vector-embeddings representing the input queries, the cosine-similarity metric was chosen to get similarity scores of sample pairs. These scores were then fed into the meta-classifier. For the meta-classification, an SVM was trained on the development split of the dataset.

5. Experiments

Table 5 displays the scores achieved by different configurations of the system on the test set.

The reciprocal rank scores achieved by each individual sentence transformer range between 0.2614 to 0.7614. The best results were achieved by the fine-tuned All-MiniLM-L6-v2 model, that was trained on the train set consisting of claims. The models that were fine-tuned on subtitles performed the worst not only on the validation set, but also

Table 6
Leaderboard of the task for the evaluation on the test set

User/team	MAP All	MAP 1	MAP 3	MAP 5	MAP 10	RR	Prec. 3	Prec. 5	Prec. 10
mshlis [11]	0.957	0.943	0.955	0.956	0.956	0.957	0.322	0.194	0.098
Viktor [12]	0.922	0.904	0.919	0.922	0.922	0.922	0.313	0.19	0.095
watheq9	0.923	0.9	0.921	0.921	0.921	0.923	0.316	0.189	0.095
Team_SimBa [13]	0.907	0.876	0.905	0.907	0.907	0.907	0.314	0.19	0.095
motlogelwan	0.878	0.833	0.87	0.873	0.876	0.878	0.306	0.187	0.095
Fraunhofer SIT	0.624	0.557	0.601	0.61	0.617	0.624	0.221	0.141	0.075
Team_Vax_Misinfo	0.096	0.005	0.011	0.02	0.054	0.096	0.006	0.011	0.033
random-baseline	0	0	0	0	0	0	0	0	0

on the test set. Thus, they were omitted from the final model.

Two types of ensemble classification techniques were experimented on during evaluation. One averages the similarity scores that are output by each weak classifier, while the other takes advantage of the proposed SVM meta-classifier. In the evaluative results it showed, that averaging the similarity scores did not enhance the performance. In fact, the performance worsened in comparison to the fine-tuned All-MiniLM-L6-v2 model. The proposed meta-classifier however was able to get the best scores when including data preprocessing and augmentation. Omitting data preprocessing and data augmentation yielded less good results, even though there were only minor differences. Unfortunately, due to an error during submission, the similarity score averaging model was submitted to the competition's leaderboard instead of the one using the trained meta-classifier. It ranked sixth in the competition (Table 6).

6. Conclusion

In this paper, a new approach to detecting whether statements in tweets were previously fact-checked. It consisted of multiple fine-tuned sentence transformers served as weak classifiers in an ensemble classification scheme. Instead of only taking the claims of the fact-check into consideration during ranking, the semantic similarity between a given tweet and the title or the subtitle of the published fact-check was analyzed as well. An SVM was then used to as a meta-classifier. Further, it took advantage of data augmentation, in particular back-translation, and data preprocessing. Experimental results have shown, that while the system was indeed able to identify which statements were subject to a previously conducted fact-checking, the system still needs some improvements in order to be used in practice. Interestingly, the ranking did not benefit from the utilization of the estimated semantic similarity between tweets and subtitles of the fact-checks during classification. Future work could revolve around including a re-ranker as well as taking advantage of a cross-encoder.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) and the and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of “ATHENE – Disinformation on Corona” and “Lernlabor Cybersicherheit”.

References

[1] G. K. Shahi, J. M. StruSS, T. Mandl, Overview of the CLEF-2021 CheckThat! lab task 3 on fake news detection, in: Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum, CLEF ’2021, Bucharest, Romania (online), 2021. URL: http://ceur-ws.org/Vol-2936/paper-30.pdf.

[2] 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, 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.

[3] 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.

[4] P. Nakov, G. Da San Martino, 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, in: Proceedings of the 12th International Conference of the CLEF Association: Information Access Evaluation Meets Multiliguality, Multimodality, and Visualization, CLEF ’2021, Bucharest, Romania (online), 2021. URL: https://link.springer.com/chapter/10.1007/978-3-030-85251-1_19.

[5] S. Shaar, F. Haouari, W. Mansour, M. Hasanain, N. Babulkov, F. Alam, G. Da San Martino, T. Elsayed, P. Nakov, Overview of the CLEF-2021 CheckThat! lab task 2 on detecting previously fact-checked claims in tweets and political debates, in: Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum, CLEF ’2021, Bucharest, Romania (online), 2021. URL: http://ceur-ws.org/Vol-2936/paper-29.pdf.

[6] A. Chernyavskiy, D. Ilvovsky, P. Nakov, Aschern at checkthat! 2021: normalcr lambda-calculus of fact-checked claims, 2021.

[7] A. Pritzkau, Nlytics at checkthat! 2021: Multi-class fake news detection of news articles and domain identification with roberta - a baseline model, in: CLEF, 2021.

[8] S. Mihaylova, I. Borisova, D. Chemishanov, P. Hadzhitsanev, M. Hardalov, P. Nakov, Dips at checkthat! 2021: verified claim retrieval, Faggioli et al.[33] (2021).

[9] J. M. Pérez, J. C. Giudici, F. Luque, pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks, 2021. arXiv:2106.09462.

[10] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[11] S. D.-H. Michael Shliselberg, RIET Lab at CheckThat! 2022: improving decoder based re-ranking for claim matching, in: N. Faggioli, Guglielmo andd Ferro, A. Hanbury, M. Potthast (Eds.), Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[12] V. Kostov, AI Rational at CheckThat! 2022: reranking previously fact-checked claims on semantic and lexical similarity, in: N. Faggioli, Guglielmo andd Ferro, A. Hanbury, M. Potthast (Eds.), Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.

[13] A. Hövelmeyer, K. Boland, S. Dietze, SimBa at CheckThat! 2022: lexical and semantic similarity based detection of verified claims in an unsupervised and supervised way, in: N. Faggioli, Guglielmo andd Ferro, A. Hanbury, M. Potthast (Eds.), Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.