

# GPLSI team at CheckThat! 2021: Fine-tuning BETO and RoBERTa

Robiert Sepúlveda-Torres, Estela Saquete

*Department of Software and Computing Systems, University of Alicante, Apdo. de Correos 99, E-03080 Alicante, Spain*

## Abstract

CheckThat! Lab is a challenging lab aimed at tackling the disinformation problem. The GPLSI team from the University of Alicante (Spain) has participated in two tasks of the CheckThat! Lab namely, Task 1 (Check-Worthiness Estimation) and Task 3 (Fake News Detection). We attained second and fifth place in the Spanish-version and English-version of Subtask 1A. Our systems use models based on transfer learning such as RoBERTa and BETO. The best results were achieved by fine-tuning these models. However, our results for Subtask 3A are low compared to the team that achieved the best result. We included some external features in the models for Subtask 1A and 3A, but we could not improve the results. In future work, we will experiment by incorporating other external features into the models with the aim of improving the results of the tasks.

## Keywords

Check-worthiness, Transfer learning models, Fake news detection

## 1. Introduction

Fake news has existed for a long time, but with the exponential rise in the consumption of news through digital media, disinformation has become one of the main problems in modern society [1]. More recently, the huge volume of news in digital media makes it impossible to evaluate its veracity manually in a reasonable time frame [2]. The scientific community is currently using artificial intelligence to address the problem by, for example, developing large-scale datasets with the aim of creating automated fact-checking systems [3].

In this context, CheckThat! Lab emerges as part of the Cross-Language Evaluation Forum (CLEF). CheckThat! Lab's goal is to foster the development of technologies that allow the automatic verification of claims [4]. This article provides a comprehensive report on the participation of the GPLSI team in Subtask 1A (Check-worthiness of tweets) and 3A (Multi-class fake news detection of news articles) of CheckThat! Lab. [4] provides a detailed description of CheckThat! Lab. The subtasks are summarized below:

1. Subtask 1A consists of predicting whether a given tweet is worth fact-checking. Subtask 1A is offered in Arabic, Bulgarian, English, Turkish, and Spanish. The GPLSI team participates in the English and Spanish version of the subtask. Subtask 1A uses Mean

---

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

✉ rsepulveda@dlsi.ua.es (R. Sepúlveda-Torres); stela@dlsi.ua.es (E. Saquete)

🆔 0000-0002-2784-2748 (R. Sepúlveda-Torres); 0000-0002-6001-5461 (E. Saquete)



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

 CEUR Workshop Proceedings (CEUR-WS.org)

Average Precision (MAP) as the official evaluation measure and we will report reciprocal rank, and P@k for  $k \in \{1, 3, 5, 10, 20, 30\}$  as well [5].

2. Subtask 3A consists of detecting fake news as a four-class classification problem. Given the title and body text of a news article, it determines whether the main claim made in the article is true, partially true, false, or other [6]. Subtask 3A is offered only in English and uses the macro F1 measure. The 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. The article contains partially true and partially false information, but cannot be considered 100% true.
  - 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.

## 2. Related Work

Automatic detection of misinformation and fake news is a complex task in which Artificial Intelligence (AI) and Natural Language Processing (NLP) play a key role. Due to the complexity of the task, different subtasks are being dealt with [7], both in traditional media [8] or social media [9]. Within these subtasks, automatic fact checking is one of the most challenging problems, and related competitions are very important as new datasets are designed to create models. In 2018, the two main competitions were: i) CheckThat! Lab, the CLEF-2018 Fact checking Lab<sup>1</sup> for the automatic identification and verification of claims in political debates[10]. The dataset delivered by this competition was obtained from FactCheck.org and it annotates statements with true/half-true/false values [11]; and, ii) the Fact Extraction and VERification (FEVER)<sup>2</sup>, which is a workshop on fact extraction and verification providing a dataset of 220K claims verified against Wikipedia [12] [13]. The statements in the corpus were annotated with supports/ refutes/ notEnoughInfo.

CheckThat! Lab 2021 is the fourth edition of the lab and Task 3 on Fake News Detection is new for this edition. Task 1 on Check-Worthiness has been present in previous editions, but it is new in the Spanish, Turkish, and Bulgarian versions. The previous year (2020) the winning team in the English-version of this task was the Accenture team [14]. This team fine-tuned the RoBERTa model and reached a MAP score of 0.8064. The second team ranked was Alex, and this team concatenated the RoBERTa model with tweet metadata [15], obtaining a MAP score of 0.8034.

There have been efforts similar to CheckThat! Lab's approach to tackling the problem of disinformation, such as [16, 17], that have developed approaches for automated fact-checking.

In recent years, the use of Transfer Learning models has become popular to tackle the main tasks within Natural Language Processing (NLP). Some of the most successful models in this

---

<sup>1</sup><http://alt.qcri.org/clef2018-factcheck/>

<sup>2</sup><http://fever.ai/>

context are BERT and RoBERTa for the English language and BETO for the Spanish language [18, 19, 20]. For example, RoBERTa has been used to predict the stance relationship between the headline and body text of an article [21].

Considering the literature, our participation in the CheckThat! Lab makes use of these Learning Transfer models to address subtasks 1A and 3A.

### **3. Neural Models**

The models used in this research are based on the BERT model. BERT is a multi-layer bidirectional Transformer encoder that is designed to pre-train from text without labels. This pre-training model has the advantage of fine-tuning capability via a single additional layer of output, a feature that facilitates the creation of state-of-the-art models in various NLP tasks [18]. For Subtask 1A and Subtask 3A in English, the RoBERTa model will be used and for Subtask 1A in Spanish, the BETO model will be used.

#### **3.1. English-version Subtask: RoBERTa model**

RoBERTa (Robustly optimized BERT approach) is a pre-training model based on BERT [19]. RoBERTa includes the following modifications: eliminating the prediction of the next sentence; performing the training on a greater volume of data; enlarging the batch size; and, lengthening the input sequence. This implementation attains state-of-the-art results in General Language Understanding Evaluation (GLUE) and Reading Comprehension Dataset From Examinations (RACE). In this research, we use RoBERTa large model architecture with 24 self-attention layers, a hidden size of 1024 and 355M parameters [19].

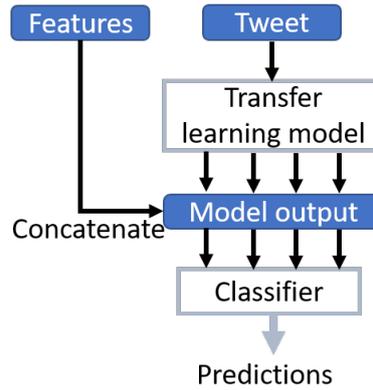
#### **3.2. Spanish-version Subtask: BETO model**

BETO also uses BERT's architecture but includes a series of optimizations similar to those performed in the RoBERTa model. The BETO model was pre-trained with Wikipedia texts and all OPUS Project sources [22] in the Spanish language. This model achieved better results in most of the tasks present in the GLUE benchmark than the multilingual models based on BERT. This model has 12 self-attention layers with a hidden size of 768 and a total of 110M parameters [20].

#### **3.3. Architecture modifications**

Some researchers include additional features to blend them with the output of the last layer of the transfer learning models [23, 24]. This strategy could improve the prediction of models based on transfer learning. In the tasks where the GPLSI team participates, we have experimented by varying our model so as to bring it closer to the domain of each task. Figure 1 shows the internal architecture of our classifier when we included external features.

In Subtask 1A, for both English and Spanish-language versions, we extract features related to the presence and quantity of numbers and dates in the tweets. The Stanza Python library is



**Figure 1:** Architecture modifications with external features

used to extract these features. Stanza library analyzes part-of-speech and morphological feature tagging, dependency parsing, and named entity recognition parser [25].

Another change that has been tested for both Subtask 1A is the inclusion of features from Linguistic Inquiry and Word Count (LIWC). LIWC is a resource for detecting meaning in a wide variety of experimental settings, including showing the focus of attention, emotionality, social relationships, thinking styles, and individual differences [26, 27]. The LIWC dictionaries have been translated into several languages, including Spanish, German, Italian and Portuguese. We used Spanish translated LIWC for the Spanish-version Subtask and the LIWC original version for the English-version Subtask.

## 4. Experiments

The objective of the experiments is to find the model and the hyperparameters that are the best possible fit for the tasks outlined. This section presents the experiments carried out based on the models described previously.

### 4.1. Subtask 1A experiments

This subtask aims to determine whether a tweet should be checked or not; hence the input to each model is a tweet and, in the case of some experiments, the input also adds external features. The input tweet to the models is pre-processed and Emoji are extracted to decrease the out-of-vocabulary words in the models being used. The proposed six experiments for this subtask use techniques that are well recognized by the scientific community, and will be described next:

1. Our baseline system with RoBERTa or BERT model: This experiment makes a fine-tuning of the corresponding model over the corpus of the task in evaluation. In line with [1] the BERT baseline system and RoBERTa baseline system have the same hyperparameters, as

follows: maximum sequence length of 125; batch size of 4; training rate of 1.5e-5; and training epochs of 3.

2. This experiment performs a Bayes search to optimize the hyperparameters. We used the Weights & Biases library to automate hyperparameter tuning and explore the space of possible models. This library enables the visualization and comparison of the results of each model [28]. The search configuration is shown in table 1.

**Table 1**

Hyperparameter tuning

Parameters	Values
Number train epochs	2, 3, or 4
Dropout	0.1, 0.2, 0.3, 0.4, or 0.5
Batch size	2, 4, or 8
Learning rate	1e-5, 1.5e-5, 2e-5, 2.5e-5, or 3e-5

3. RoBERTa or BERT best model with number and date indicators: Concatenated the output of the last layer of RoBERTa or BERT with the number and date indicators.
4. RoBERTa or BERT best model with LIWC features: Concatenated the output of the last layer of RoBERTa or BERT with the LIWC features.
5. RoBERTa or BERT best model with oversampling: The training set is extended with examples of the least representative classes to balance the dataset.
6. RoBERTa or BERT best model with undersampling: Examples of the most representative classes are eliminated to balance the dataset.

## 4.2. Subtask 3A experiments

The first two experiments of the previous subtask are used in the same way for this subtask. Subtask 3A is conducted in English and therefore only the RoBERTa model can be used. In addition, as this task classifies news according to its veracity, the RoBERTa model processes both the title and the body text of the article, representing them as a sequence of concatenated words. A third experiment is added that tries to improve the results of the two previous experiments. This experiment splits the four-class classification into two binary classifications and one three-class classification.

## 5. Result and Discussion

The experiments were implemented using the Simple Transformer<sup>3</sup> and PyTorch<sup>4</sup> libraries. The experiments are trained with the training set and their performance is evaluated with the development set provided by the organizers of CheckThat! Lab. In our experiments, the output of BERT and RoBERTa models passes through a series of layers to finish classifying the tweet or the news. The layers used, in order, are two Dropouts, and Linear. After the first Dropout layer,

<sup>3</sup><https://simpletransformers.ai/> (accessed on 20 May 2021)

<sup>4</sup><https://pytorch.org/> (accessed on 20 May 2021)

we used the Tanh activation function. The tweet classification is a binary classification so the output layer has a single neuron and binary cross-entropy is used as the loss function. However, the news is classified into four classes, so the output layer has 4 neurons and cross-entropy is used as the loss function.

## 5.1. Subtask 1A

The Spanish-version of Subtask 1A has a dataset of 2,495 tweets for training, 1,247 for development, and 1248 for testing. The English-version of Subtask 1A has a dataset of 822 tweets for training, 140 for development and 350 for testing.

The BETO baseline system obtains good results in the macro F1 metric; however, the evaluation metric of this subtask is MAP and in this case, the results are quite close to the baseline provided by the organizers. Similarly to the behavior of the BETO baseline system, the RoBERTa baseline system achieves good results on the macro F1 metric. However, the MAP metric is 4 points lower than 0.8064, which was reached by the best competitor of CheckThat! Lab 2020.

Experiment 2 is the experiment with the best results in both languages. With the help of the Weights & Biases library, a configuration has been found, although it cannot be guaranteed that it is the best one given that the search was not exhaustive. In the Spanish-version, the hyperparameter configurations are a maximum sequence length of 125, batch size of 8, training rate of 1e-5, dropout rate of 0.2, and, training performed for 2 epochs. For the English-version, the hyperparameter configurations are as follows: maximum sequence length of 125; batch size of 4; training rate of 1.5e-5; dropout rate of 0.2; training performed for 3 epochs. Table 2 shows the results.

**Table 2**

CheckThat! Spanish-version and English-version of Subtask 1A experiments with BETO and RoBERTa models in the development dataset

No	Experiment	Spanish		English	
		MAP	Macro F1	MAP	Macro F1
1	BETO and RoBERTa baseline system	0.485	0.709	0.762	0.702
2	Hyperparameter tuning	<b>0.549</b>	<b>0.712</b>	<b>0.825</b>	<b>0.750</b>
3	Best model with number and date indicators	0.500	0.633	0.652	0.706
4	Best model with LIWC features	0.497	0.685	0.624	0.624
5	Best model with oversampling	0.387	0.693	0.772	0.735
6	Best model with undersampling	0.455	0.564	0.795	0.709

Experiment 3 indicates that the features obtained do not help to identify whether a tweet should be checked or not. Experiment 4 worsened the results for both languages. Experiments 5 and 6 also did not improve the results achieved by simply fine-tuning the basic model. In both cases, the two metrics descended in relation to experiment 2.

The GPLSI team reached second place in the Spanish-version of Subtask 1A to predict the test set. The difference in the MAP metric is less than one point with respect to the team that came first in this subtask. Table 3 shows the results.

In the English-version of Subtask 1A, the GPLSI team reached fifth place. In this case,

**Table 3**  
CheckThat! Spanish-version of Subtask 1A results

Team	MAP	MRR	RP	P@1	P@3	P@5	P@10	P@20	P@30
GPLSI	0.529	0.500	0.533	0.000	0.667	0.600	0.800	0.750	0.620

the difference between the first-place team and GPLSI was greater than that observed in the Spanish-version of the same subtask. Table 4 shows the results.

**Table 4**  
CheckThat! English-version Subtask 1A result

Team	MAP	MRR	RP	P@1	P@3	P@5	P@10	P@20	P@30
GPLSI	0.132	0.167	0.158	0.000	0.000	0.000	0.200	0.150	0.140

To sum up, the numerous experiments that were conducted failed to improve the results achieved by experiment 2, indicating the power of models based on transfer learning for this task. Evidently, we have not been able to find appropriate external features to tackle this subtask. The systems developed in experiment 2 are used to predict the test set for both languages.

## 5.2. Subtask 3A

The training dataset available for this subtask contains 900 news items [29]. In order to evaluate the models that were being fine-tuned, the training dataset was divided into a training set and a development set. The split proportion is of 0.7 and 0.3 for a new training set and a development set. The training and development sets maintain similar percentages of examples from each class. The test set has 365 news stories.

Our baseline system uses the hyperparameters described in experiment 1 of the previous section by only changing the maximum sequence length to 512. Task 3A is considered more complicated than 1A because it is necessary to find patterns that classify the news into 4 classes. The macro F1 result of the baselines fine-tuning RoBERTa is quite low which corroborates the complexity of this subtask.

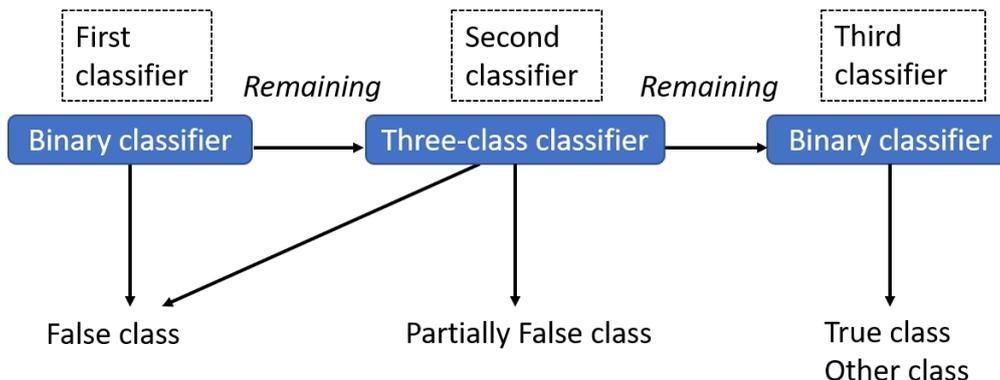
Experiment 2 carried out a deep hyperparameter tuning and, as the results show, the improvement is negligible. In another task, a hyperparameter tuning like this should have improved the baseline significantly.

**Table 5**  
CheckThat! Subtask 3 English experiments with RoBERTa model in development dataset

No	Experiment	Macro F1
1	Our baseline system	0.516
2	Hyperparameter tuning	0.520
3	Best model with three classifiers	0.548

The last experiment is depicted in figure 2. The strategy involved placing the majority class in the first classifier and the minority classes in the subsequent classifiers. In the first classifier, we

predict False and Remaining classes (Partially False, True, and Other). In the second classifier, we predict False, Partially False and Remaining classes (True and Other). Finally, the third classifier is True and Other. Each classifier specializes in predicting a group of classes and the remaining classes are passed to subsequent classifiers.



**Figure 2:** Explanation of the classification pipeline

- The first classifier achieves 0.802 as macro F1 with hyperparameter configurations are as follows: maximum sequence length of 512; batch size of 2; training rate of  $2e-5$ ; dropout rate of 0.2; training performed for 3 epochs.
- The second classifier achieves 0.666 as macro F1 with hyperparameter configurations are as follows: maximum sequence length of 512; batch size of 2; training rate of  $2e-5$ ; dropout rate of 0.2; training performed for 4 epochs.
- The third classifier achieves 0.727 as macro F1 with hyperparameter configurations are as follows: maximum sequence length of 512; batch size of 4; training rate of  $1e-5$ ; dropout rate of 0.2 training performed for 3 epochs.

The GPLSI team ranked 16th in this subtask. In this subtask, we failed to find a competitive model that could obtain state-of-the-art results.

## 6. Conclusions

We participated in two tasks of the three CheckThat! tasks. The results achieved in the Spanish-version of subtask 1A are considered good. They confirm that, depending on the task, fine-tuning pre-trained models may be a good option. In the Spanish-version, we ranked second with a MAP score of 0.529 and in the English-version we ranked fifth with a MAP score of 0.132.

On the other hand, the results obtained in Subtask 3A leave considerable room for improvement and to date, no enhancement was found in the models used. However, the classifier cascade technique improved the classification. In this research, we classify the most majority classes in the first classifiers. We include some external features to the models used but the results obtained do not improve the fine-tuning experiment of each model.

In future work, we will experiment with other reference neural models and look for specific features that can improve the results for the most complicated tasks.

## Acknowledgments

This research work has been partially funded by Generalitat Valenciana through project “SIIA: Tecnologías del lenguaje humano para una sociedad inclusiva, igualitaria, y accesible” (PROME-TEU/2018/089), by the Spanish Government through project “Modelang: Modeling the behavior of digital entities by Human Language Technologies” (RTI2018-094653-B-C22), and project “INTEGER - Intelligent Text Generation” (RTI2018-094649-B-I00). Also, this paper is also based upon work from COST Action CA18231 “Multi3Generation: Multi-task, Multilingual, Multi-modal Language Generation”.

## References

- [1] R. Sepúlveda-Torres, A. Bonet-Jover, E. Saquete, “Here Are the Rules: Ignore All Rules”: Automatic Contradiction Detection in Spanish, *Applied Sciences* 11 (2021) 3060.
- [2] G. Tsipursky, F. Votta, K. M. Roose, Fighting Fake News and Post-Truth Politics with Behavioral Science: The Pro-Truth Pledge, *Behavior and Social Issues* 27 (2018) 47–70.
- [3] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and verification, *arXiv preprint arXiv:1803.05355* (2018) 809–819.
- [4] P. Nakov, G. D. S. Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, The CLEF-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: D. Hiemstra, M. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*, volume 12657 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 639–649.
- [5] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, M. K. Alex Nikolov, F. A. Yavuz Selim Kartal, 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: *Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum, CLEF ’2021, Bucharest, Romania (online)*, 2021.
- [6] 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.
- [7] E. Saquete, D. Tomás, P. Moreda, P. Martínez-Barco, M. Palomar, Fighting post-truth using natural language processing: A review and open challenges, *Expert Systems with Applications* 141 (2020) 112943.
- [8] A. Bonet-Jover, A. Piad-Morffis, E. Saquete, P. Martínez-Barco, M. Ángel García-Cumbreras, Exploiting discourse structure of traditional digital media to enhance automatic fake news detection, *Expert Systems with Applications* (2020) 114340.

- [9] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake News Detection on Social Media, *ACM SIGKDD Explorations Newsletter* 19 (2017) 22–36.
- [10] P. Nakov, L. Màrquez, A. Barrón-Cedeño, W. Zaghouani, T. Elsayed, R. Suwaileh, P. Gencheva, CLEF-2018 lab on automatic identification and verification of claims in political debates, in: *Proceedings of the CLEF-2018*, 2018.
- [11] A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, P. Atanasova, W. Zaghouani, S. Kyuchukov, G. Da San Martino, P. Nakov, Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims, task 2: Factuality, in: L. Cappellato, N. Ferro, J.-Y. Nie, L. Soulier (Eds.), *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, CEUR-WS.org, Avignon, France, 2018.
- [12] 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, 2018, pp. 1–9. URL: <http://aclweb.org/anthology/W18-5501>.
- [13] 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, 2018, pp. 809–819. URL: <http://aclweb.org/anthology/N18-1074>. doi:10.18653/v1/N18-1074.
- [14] E. Williams, Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models, Technical Report, 2020.
- [15] A. Nikolov, G. Da, S. Martino, I. Koychev, P. Nakov, Team Alex at CLEF CheckThat! 2020: Identifying Check-Worthy Tweets With Transformer Models, Technical Report, 2020.
- [16] B. S. Andreas Hanselowski, Avinesh PVS, F. Caspelherr, Description of the system developed by team athene in the FNC-1, 2017.
- [17] A. Alonso-Reina, R. Sepúlveda-Torres, E. Saquete, M. Palomar, Team gplsi. approach for automated fact checking, in: *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, 2019, pp. 110–114.
- [18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018). [arXiv:1810.04805](https://arxiv.org/abs/1810.04805).
- [19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019). [arXiv:1907.11692](https://arxiv.org/abs/1907.11692).
- [20] J. Canete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained bert model and evaluation data, PML4DC at ICLR 2020 (2020).
- [21] R. Sepúlveda-Torres, M. Vicente, E. Saquete, E. Lloret, M. Palomar, Exploring summarization to enhance headline stance detection, in: *International Conference on Applications of Natural Language to Information Systems*, Springer, 2021, pp. 243–254.
- [22] J. Tiedemann, Parallel data, tools and interfaces in OPUS, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 2214–2218.
- [23] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, X. Zhou, Semantics-aware BERT for Language Understanding, Technical Report, 2020.

- [24] W. M. Lim, H. T. Madabushi, UoB at SemEval-2020 Task 12: Boosting BERT with Corpus Level Information (2020).
- [25] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A Python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020.
- [26] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn, The development and psychometric properties of LIWC2015, Technical Report, 2015.
- [27] Y. R. Tausczik, J. W. Pennebaker, The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods, *Journal of Language and Social Psychology* 29 (2010) 24–54.
- [28] L. Biewald, Experiment tracking with weights and biases, 2020. URL: <https://www.wandb.com/>, software available from wandb.com.
- [29] G. K. Shahi, J. M. Struß, T. Mandl, Task 3: Fake news detection at CLEF-2021 CheckThat!, 2021. URL: <https://doi.org/10.5281/zenodo.4714517>. doi:10.5281/zenodo.4714517.