Accenture at CheckThat! 2023: Identifying Claims with Societal Impact using NLP Data Augmentation

Sieu Tran[1], Paul Rodrigues[1], Benjamin Strauss[1] and Evan M. Williams[2]

[1]Accenture, 1201 New York Ave NW, Washington, DC 20005, United States
[2]Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States

Abstract

This paper discusses the results of the Accenture Team in the 2023 CheckThat Lab! focusing on check-worthiness classification on multigenre text. Check-worthiness classification is similar to misinformation detection or claim detection, but informed by the potential societal impact of the claims. We utilized high quality back-translation to augment the minority classes in labeled English, Arabic, and Spanish datasets and fine-tune pre-trained foundation models for each of the languages. This method placed 2nd in Arabic, 3rd in English, and 5th in Spanish. We further show that high-quality translation is preferable for data augmentation to translation with lower BLEU scores, and that using NLP data augmentation to increase the minority class in quantities over the minority class shows promise on this task.

Keywords

data augmentation, check-worthy detection, misinformation detection, claim detection,

1. Introduction

Fact-checkers and journalists must continually evaluate their information environment and make determinations of which claims are most worthy of their effort to verify and disseminate. As CLEF's CheckThat! labs have shown over the last several years, this task is challenging. In previous iterations of this lab, annotators were asked to label check-worthiness using the three following criterion [1, 2]:

- Do you think the claim in the text is of interest to the public?
- To what extent do you think the claim can negatively affect the reputation of an entity, country, etc.?
- Do you think journalists will be interested in covering the spread of the claim or the information discussed by the claim?

The claim "all leopards are pink" is easily falsifiable, but annotators would likely agree that it does not meet any of the three criterion above and should therefore not be considered check-worthy. In contrast to misinformation detection or claim detection tasks, in order to approximate the

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18-21, 2023, Thessaloniki, Greece
sieu.tran@accenturefederal.com (S. Tran); paul.rodrigues@accenturefederal.com (P. Rodrigues); b.strauss@accenturefederal.com (B. Strauss); emwillia@andrew.cmu.edu (E. M. Williams)
0000-0003-0017-4329 (S. Tran); 0000-0002-2151-636X (P. Rodrigues); 0000-0002-0224-424X (B. Strauss); 0000-0002-0534-9450 (E. M. Williams)

Table 1
Dataset descriptions

Language	Modeling set	# of samples	Unique word count
Arabic	Train	6,059	40,539
	Test	500	6,574
	Validation	1,274	12,775
English	Train	16,876	10,326
	Test	318	980
	Validation	5,625	6,602
Spanish	Train	7,488	33,400
	Test	5,000	18,323
	Validation	2,460	14,444

function that generated check-worthiness labels, a successful model must attend to signals in text that suggest both public interest and societal impact.

Methods for automated check-worthiness identification have been evaluated for years across numerous languages by CLEF's CheckThat! Lab [2, 3, 4]. This task is particularly challenging due to the lack of context around individual observations. For example, in the gold test-set labels of this year's challenge, "He has never offered a plan" was labeled as check-worthy, while "he's been a professor for a long time at a great school" is not. In both cases, without any additional context, the subject is unknown, while the predicate in the first case hints at a subject that would be of interest to the public.

Additionally, mimicking reality, where the vast majority of content one might be exposed to online or on social media is not worthy of distribution and therefore not check-worthy, class imbalance has been a common feature of this task. In previous years, our team experimented with back-translation [5] and contextually-sensitive augmentation [6] to rebalance classes. Back-translation uses synthetic data, translated from source data through an intermediary language and then back to the source language to amplify training data. [7] Our back-translation-augmented performance was more consistent across language tasks than our contextually-sensitive augmentation performance, so we elect to exclusively use back-translation augmentation in the 2023 CheckThat! lab.

In this work, we describe the methodology of our 2023 CheckThat 1B submissions for Arabic, English, and Spanish language submissions. Task 1B focused on checkworthiness detection for multigenre language data. Our submission resulted in the 3rd highest F1 score in English, the 2nd highest F1 score in Arabic, and the 5th highest F1 score in Spanish.

2. Exploratory Analysis

Table 1 shows the number of samples and unique tokens for each of the datasets provided. We see that English had the largest number of samples in training (16,876) while Arabic had the least (6,059). However, Arabic had the highest count of unique words (40,539), due to its morphological structure, and English had the lowest (10,326).

Table 2
Unknown token distribution in data for each language.

Language	Tokenizer Type	Modeling Set	WordPiece	Unknown Token	Unknown Percent (%)
Arabic	BERT	Training	324,625	985	0.30
		Testing	28,702	163	0.57
		Validation	68,987	184	0.27
English	RoBERTa-based	Training	357,526	0	0
		Testing	5,013	0	0
		Validation	117,319	0	0
Spanish	RoBERTa-based	Training	476,395	0	0
		Testing	192,471	0	0
		Validation	140,179	0	0

2.1. Label Balance

Each of the datasets provided by the CheckThat! organizers had label bias which skewed the datasets towards texts that were not considered check-worthy. The Spanish dataset had the highest percentage of check-worthy texts (29%), followed by Arabic (29%), and then English (24%).

2.2. WordPiece Analysis

Transformer models utilize WordPiece tokenization schemes that differ and are dependant on the model. At the time of pre-training, the WordPiece algorithm determines which pieces of words will be retained, and which will be discarded. An "unknown" (UNK) token is utilized as a placeholder in the lexicon, and used to represent WordPiece tokens received in novel input that did not get utilized at model creation. We expect language samples which have a high amount of tokens processed as UNK would perform poorly.

We present our analysis in Table 2. Most notably, Spanish training set contains over 470K WordPieces, the largest number across all three languages, second by just over 350K for English. In addition, Arabic training set produced a low rate of unknown tokens (0.30%). Unexpectedly, the RoBERTa tokenizers we used did not return UNK tokens on any dataset provided by the CLEF CheckThat! organizers.

3. Transformer Architectures and Pre-Trained Models

In this work, we utilize BERT and RoBERTa models. The Bidirectional Encoder Representation Transformer (BERT) is a transformer-based architecture that was introduced in 2018 [8]. BERT has had a substantial impact on the field of NLP, and achieved state of the art results on 11 NLP benchmarks at the time of its release. RoBERTa, introduced by [9], modified various parts of BERTs training process. These modifications include more training data, more pre-training steps with bigger batches over more data, removing BERT's Next Sentence Prediction, training on longer sequences, and dynamically changing the masking pattern applied to the training

data [9].

For the Arabic Dataset, we used *lanwuwei/GigaBERT-v4-Arabic-and-English* [10], which was trained on a large-scale corpus (Arabic version of OSCAR, an Arabic Wikipedia dump, and Gigaword) with ∼10B tokens. The model showing state-of-the-art zero-shot transfer performance from English to Arabic on information extraction tasks. The Arabic model contains a vocabulary of length ∼21,000 and the English model has a vocabulary length of ∼26,000. For Spanish, we used *bertin-project/bertin-roberta-base-spanish* [11]. The Spanish RoBERTa model contains a vocabulary of length 50,261.

For English, we used *roberta-large* [9]. The English RoBERTa model contains 50,265 Word-Pieces.

4. Method

4.1. Data Augmentation

The organizers provided a training and a development set for each language. We use the provided training set and development set to create internal training and validation sets for experimentation. We use the test set provided by organizers as a hold-out test set.

For each language, augmentation and training were done with via back-translation using AWS translation. We appended back-translated check-worthy texts to the training set. In our 2021 experiment [6], we found that this form of augmentation resulted in a significant increase in recall and F1-score for check-worthy texts. For Arabic and Spanish, we used English as the pivot language which has demonstrated success in previous CheckThat Labs [5, 6]. For the English training set, due to significant sample imbalance, we augmented the positive-label data twice: the first using Arabic as the pivot language (i.e., English > Arabic > English) and the second using Arabic and Spanish as pivots (i.e., English > Arabic > English > Spanish > English).

In this work, we experiment with different quality of translation to observe how quality of augmented data improve the final model performance. Due to limited time and resources, we focused our translation experimentation on the Arabic dataset. We use the open-source translation models *helsinki-nlp/opus-mt-ar-en* and *helsinki-nlp/opus-mt-en-ar* [12] to translate to and from English, respectively.

Table 3 shows the BLEU score for each back-translation scheme. The higher the score, the more consistent or similar the translation to the original text. Arabic (0.3895) and the second back-translation for English (0.3400) show the highest level of divergence from the original text. We hypothesize this leads to more diverse data and better performing models.

Table 4 shows the number of unique tokens in the source data, the number of unique tokens in the translated augmentation data, and the difference-the number of unique tokens that was added in the translated data that was not originally in the source data. Machine translation added 6315 novel tokens to our Arabic training data, 2743 to our English training data, and 5208 to our Spanish training data.

Table 3
Average Sentence BLEU Score for Each Back-translation Scheme

Language	Back-translation	Average Sentence BLEU Score
Arabic	AR > EN > AR	0.390
English	EN > AR > EN	0.455
English	EN > AR > EN > ES > EN	0.340
Spanish	ES > EN > ES	0.551

Table 4
New Tokens in Machine Translated Text

Language	Back-translation	Unique tokens in source	Unique tokens in MT	New Tokens in MT
Arabic	AR > EN > AR	16132	14491	6315
English	EN > AR > EN	9621	9154	2252
English	EN > AR > EN > ES > EN	9621	9335	2743
Spanish	ES > EN > ES	20524	19450	5208

4.2. Classification

For both BERT and RoBERTa, we added an additional mean-pooling layer and dropout layer on top of the model prior to the final classification layer. Adding these additional layers has been shown to help prevent over-fitting while fine-tuning. We used an Adam optimizer with a learning rate of $2e-5$ and an epsilon of $1.5e-8$. We use a binary cross-entropy loss function, 4 epochs, and a batch size of 32.

5. Results

Table 5 contains all model performance on the test set provided by the organizers. We received a weighted average F1-score of 0.687 for Arabic, 0.910 for English, and 0.912 for Spanish. The official scoring of the shared task had 0.733 for Arabic yielding 2nd place, 0.860 for English yielding 3rd place, and 0.509 for Spanish yielding 5th place.

Table 6 contains Arabic model performance with various quality and quantity of back-translation augmentation on the gold test set. We received a weighted average F1-score of 0.600 with no augmentation and a 0.601 with HelsinkiNLP back-translation, showing very little aggregate improvement with this translation system. AWS back-translation provided a weighted average F1-score of 0.687, showing that higher quality back-translation provides better classification results downstream. Combining AWS and Helsinki back-translation provided a score of 0.727, showing quantity of samples (increasing quantity of the initial minority class over the majority class) increases performance as well.

Table 5
Accenture results from CheckThat! 2023 Task 1

Language	Class	Precision	Recall	F1-score
Arabic	No	0.409	0.821	0.546
	Yes	0.913	0.613	0.733
	macro avg	0.661	0.717	0.640
	weighted avg	0.789	0.664	0.687
English	No	0.903	0.971	0.936
	Yes	0.935	0.796	0.860
	macro avg	0.919	0.884	0.898
	weighted avg	0.914	0.912	0.910
Spanish	No	0.935	0.982	0.958
	Yes	0.715	0.395	0.509
	macro avg	0.825	0.689	0.733
	weighted avg	0.912	0.922	0.912

Language	Accuracy
Arabic	0.664
English	0.912
Spanish	0.922

6. Conclusion

This paper discussed the results of the Accenture team in the 2023 CheckThat Lab, Task 1B, focused on labeling the check-worthiness of Arabic, English, and Spanish multi-genre content. We utilized high quality back-translation as a method of training data augmentation for the tweets, debates, and transcripts in the challenge and placed 2nd in Arabic, 3rd in English, and 5th in Spanish. We showed, on Arabic, that a better performing translation system improves performance in the downstream task. Additionally, we showed that a strategy of rebalancing the training data, by using NLP data augmentation to flip the minority class to the positive class may be beneficial in this task.

References

[1] A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, K. Eickhoff, A. Névéol, L. Cappellato, N. Ferro, Experimental ir meets multilinguality, multimodality, and interaction, in: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020), volume 12260, Springer, 2020.

[2] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, et al., Overview of CheckThat! 2020: Automatic identification and verification of claims in social media, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the

Table 6
Accenture results from experimentation with different translation quality to improve BERT-based Arabic model

Augmentation	Class	Training Sample	Precision	Recall	F1-score
No augmentation	No	4,301	0.363	0.935	0.523
	Yes	1,758	0.956	0.464	0.620
	macro avg		0.660	0.700	0.574
	weighted avg		0.810	0.580	0.600
Back-translation with HelinskiNLP	No	4,301	0.347	0.821	0.488
	Yes	3,516	0.895	0.496	0.638
	macro avg		0.621	0.659	0.563
	weighted avg		0.760	0.576	0.601
Back-translation with AWS translation	No	4,301	0.409	0.821	0.546
	Yes	3,516	0.913	0.613	0.733
	macro avg		0.661	0.717	0.640
	weighted avg		0.789	0.664	0.687
Back-translation with AWS translation and HelinskiNLP	No	4,301	0.447	0.780	0.568
	Yes	5,274	0.905	0.684	0.779
	macro avg		0.676	0.732	0.674
	weighted avg		0.792	0.708	0.727

Augmentation	Accuracy
No augmentation	0.580
Back-translation with HelinskiNLP	0.576
Back-translation with AWS translation	0.664
Back-translation with AWS translation and HelinskiNLP	0.708

CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11, Springer, 2020, pp. 215–236.

[3] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, et al., Overview of the CLEF–2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12, Springer, 2021, pp. 264–291.

[4] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets (2022).

[5] E. Williams, P. Rodrigues, V. Novak, Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models, arXiv preprint arXiv:2009.02431 (2020).

[6] E. Williams, P. Rodrigues, S. Tran, Accenture at CheckThat! 2021: interesting claim identification and ranking with contextually sensitive lexical training data augmentation,

arXiv preprint arXiv:2107.05684 (2021).

[7] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 86–96. URL: https://aclanthology.org/P16-1009. doi:10.18653/v1/P16-1009.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[9] 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, arXiv preprint arXiv:1907.11692 (2019).

[10] W. Lan, Y. Chen, W. Xu, A. Ritter, An empirical study of pre-trained transformers for Arabic information extraction, arXiv preprint arXiv:2004.14519 (2020).

[11] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, BERTIN: Efficient pre-training of a Spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.

[12] J. Tiedemann, S. Thottingal, OPUS-MT — Building open translation services for the World, in: Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT), Lisbon, Portugal, 2020.