Accenture at CheckThat! 2023: Impacts of Back-translation on Subjectivity Detection

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 CLEF CheckThat! Lab Task 2 on Subjectivity in News Articles, and our approach on using back-translation to augment the minority classes in Arabic, English, Turkish, German, Italian, and Dutch to distinguish subjective and objective statements. While we find that back-translation works well for other tasks in the fact-checking pipeline, we find that it does not work as well for subjectivity detection. This paper begins to examine several reasons why back-translation as an NLP data augmentation strategy could inhibit subjectivity detection.

Keywords
subjectivity detection, opinion detection, news analysis, data-driven journalism

1. Introduction

Subjectivity detection, a subtask of sentiment analysis, aims to differentiate neutral content or facts from opinion within text [1]. As sentiment analysis is often concerned with the opinions of users, the removal of neutral or objective text is a common pre-processing step, particularly in polarity-detection settings [2]. However, recent work has explored the usefulness of subjectivity detection systems outside sentiment-oriented tasks, such as in augmenting fake news detection systems [3, 4, 5]. [4] use subjectivity lexicons to help differentiate and classify real and fake news in English and Brazilian Portuguese, but found that simpler BOW methods outperformed their lexicons. [3] perform statistical analyses to demonstrate a relationship between subjective language and fake news. [5] demonstrated that fine-tuned transformer-base models can perform very well on sentence-level subjectivity detection tasks.

Building on these new developments, Task 2 of the CheckThat! Lab at CLEF 2023 provides participants with annotated news sentence subjectivity detection datasets in Arabic, English, Turkish, German, Italian, and Dutch [6]. In news articles, particularly in biased settings, subjectivity detection and annotation is a challenging task, as sentences can contain both objective claims and subjective framing. For example, in the English validation dataset for Task 2, the sentence, "Wing is also the co-author of several Leftist indoctrination books for

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)

children, including one entitled What Is White Privilege?" is labeled as 'Objective', rather than 'Subjective'. As the sentence contains specific, falsifiable claims, this seems to be a reasonable labeling. However, the characterization of the books as tools of 'Leftist indoctrination', is clearly a subjective editorialization on the part of the author. This highlights the inherent ambiguity present in the task and underscores a core challenge that the annotators, and the models both face in learning a clear decision boundary.

In this work, we describe the back-translation augmentation strategies and models employed by Team Accenture's submissions to Task 2. Team Accenture's back-translation and transformer approach yielded the 3rd highest submissions in Arabic, 4th in Turkish, 5th in Dutch, and 8th in German and English. While back-translation has been shown to be an effective means of NLP data augmentation to improve checkworthiness identification [7], we speculate that the approach may reduce the the ability of models to generalize in a subjectivity detection task and explore some reasons why this may be the case.

2. Exploratory Analysis

Table 1 shows the number of samples and unique word counts for each of the datasets provided. We see that Italian had the largest number of samples in training (1,613). However, Arabic had the highest count of unique words (12,181), while German (4,622) and Dutch (3,944) had the lowest. Assuming consistent data collection methodology and annotation standards across languages, we would hypothesize that a larger quantity of unique words would yield higher-accuracy models. The sample size of all languages in this task is relatively small compared to the other tasks in the CheckThat Lab.

As shown in Figure 1, all of the datasets provided by the CheckThat! organizers had label bias which skewed each dataset towards sentences labeled as 'objective'.

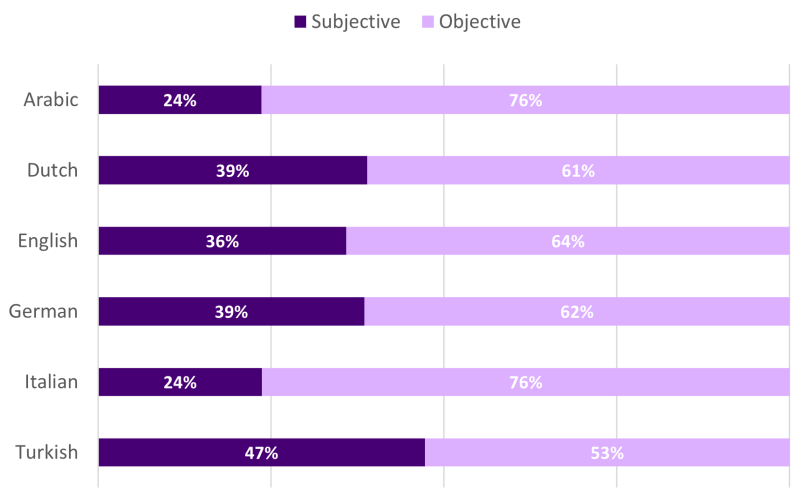

Figure 1: Label distribution across training sets

Table 1
Dataset Descriptions

Language	Modeling set	# of samples	Unique word count
Arabic	Train	1,185	12,181
	Test	445	6,225
	Validation	297	4,631
Dutch	Train	800	3,944
	Test	500	2,615
	Validation	200	1,462
English	Train	830	4,126
	Test	243	2,043
	Validation	219	1,846
German	Train	800	4,622
	Test	291	2,384
	Validation	200	1,633
Italian	Train	1,613	7,372
	Test	440	3,563
	Validation	227	1,649
Turkish	Train	800	4,914
	Test	240	1,886
	Validation	200	1,624

Transformer models utilize WordPiece tokenization schemes that are dependant on the model being evaluated. 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.

The proportion of out-of-vocabulary tokens are have been shown to inversely correlates to overall accuracy [8], so we explore proportions of UNK in each dataset to ensure our models are not excluding too many tokens from any language. We present our analysis in Table 2. Most notably, Arabic training set has the highest WordPiece count of 43,601. Since the unknown token rates are mostly negligible between all languages, we expect count and diversity of Wordpiece would influence model performance the most. 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 [9]. 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 [10], 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

Table 2
Unknown Token Distribution in Data for Each Language.

Language	Tokenizer Type	Modeling Set	WordPiece	Unknown Token
Arabic	BERT-based	Training	43,601	3
		Testing	16,050	8
		Validation	11,286	3
Dutch	BERT-based	Training	19,033	3
		Testing	10,997	0
		Validation	4,902	0
English	RoBERTa-based	Training	24,147	0
		Testing	7,674	0
		Validation	6,935	0
German	BERT-based	Training	21,318	7
		Testing	8,293	7
		Validation	5,267	4
Italian	BERT-based	Training	41,767	2
		Testing	14,978	2
		Validation	5,277	0
Turkish	BERT-based	Training	16,593	5
		Testing	4,795	4
		Validation	4,008	2

data [10].

For the Arabic Dataset, we used *lanwuwei/GigaBERT-v4-Arabic-and-English* [11], which was trained on a large-scale corpus (Arabic version of OSCAR, an Arabic Wikipedia dump, and Gigaword) with $\sim$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 $\sim$21,000 and $\sim$26,000 for English and Arabic respectively.
For English, we used *roberta-large* [10]. The English RoBERTa model contains 50,265 WordPieces. For Turkish, German, and Italian, we used *dbmdz/bert-base-turkish-cased* [12], *dbmdz/bert-base-german-uncased* [13], and *dbmdz/bert-base-italian-xxl-uncased* [14], respectively. The vocabulary sizes of the Turkish, German, and Italian models are respectively 32,000, 31,102, and 32,102. For Dutch, we used *GroNLP/bert-base-dutch-cased* [15], which has a vocabulary size of 30,073. The foundation model for each language was selected based on models we have used in the past. Recognizing that this was a problem that should not benefit from case signaling, we chose the uncased variant for any new model.

For experimentation and comparison to *roberta-large*, we also fine-tune the pre-trained model on subjectivity/style classification task, *cffl/bert-base-styleclassification-subjective-neutral* [16]. This BERT-based model has been fine-tuned on the Wiki Neutrality Corpus (WNC) - a parallel corpus of 180,000 biased and neutralized sentence pairs along with contextual sentences and metadata. The model can be used to classify text as subjectively biased vs. neutrally toned.

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

Language	Back-translation	Average Sentence BLEU Score
Arabic	AR > EN > AR	0.224
	AR > EN > ES > EN > AR	0.156
	AR > EN > FR > EN > AR	0.135
Dutch	NL > EN > NL	0.434
English	EN > ES > EN	0.428
German	DE > EN > DE	0.357
Italian	IT > EN > IT	0.456
	IT > EN > ES > EN > IT	0.353
	IT > EN > FR > EN > IT	0.313
Turkish	TR > EN > TR	0.105

4. Method

4.1. Data Augmentation

For each language, augmentation and training were done via back-translation into the respective language using AWS translation. We back-translated the minority class in each dataset, which is always the subjective documents. We appended back-translated subjective documents to the training set. In our 2021 experiment [7], we found that this form of augmentation resulted in a significant increase in recall and F1-score for the positive class. We did not use any dataset outside the one provided by the organizers for data augmentation.

In this work, we fine-tune *lanwuwei/GigaBERT-v4-Arabic-and-English* at different levels of data augmentation and compare performances on the gold test set provided by the organizer.

Table 3 shows the BLEU score for each back-translation scheme. Table 4 show training sample size before and after data augmentation and Table 5 shows the number of new tokens acquired after back-translation for each language. The higher the score, the more consistent or similar the translation to the original text. For Arabic and Italian, BLEU scores decrease as more pivot languages are used for back-translation, as we would expect. As a perfect translation would not provide variation in the training samples, and a low BLEU score may not provide consistent variation, this may suggest there is a sweet spot to BLEU score in a NLP data augmentation task to provide diverse word selection but consistent translations.

4.2. Classification

For all BERT and RoBERTa models utilized across all languages, 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.

Table 4
Training Sample Size Before and After Data Augmentation

Language	Label	Orginial Dataset Sample Count	Augmented Dataset Sample Count
Arabic	SUBJ	280	840
	OBJ	905	905
Dutch	SUBJ	311	622
	OBJ	489	489
English	SUBJ	298	596
	OBJ	532	532
German	SUBJ	308	616
	OBJ	492	492
Italian	SUBJ	382	1146
	OBJ	1231	1231
Turkish	SUBJ	378	756
	OBJ	422	422

Table 5
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	4717	4384	2166
	AR > EN > ES > EN > AR	4717	4361	2456
	AR > EN > FR > EN > AR	4717	4373	2541
Dutch	NL > EN > NL	2406	2323	732
English	EN > ES > EN	2590	2527	787
German	DE > EN > DE	2432	2361	808
Italian	IT > EN > IT	3309	3209	928
	IT > EN > ES > EN > IT	3309	3199	1134
	IT > EN > FR > EN > IT	3309	3206	1238
Turkish	TR > EN > TR	2967	2813	1533

5. Results

Table 6 and 7 contains all model performance on the test set provided by the organizers. We find that our Arabic model has an accuracy of 0.800 with a weighted average F1-score of 0.816. Our English model had an accuracy of 0.696 with a weighted average F1-score of 0.687. For Turkish, we had an accuracy of 0.788 and a weighted average F1-score of 0.784. German received an accuracy of 0.337 and an F1-score of 0.174. Italian had an accuracy of 0.689 and F1 of 0.706. Finally, our Dutch model had an accuracy of 0.646 and a weighted F1-score of 0.618.

Table 8 and 9 shows Arabic model's performance on the gold test set with different level of data augmentation.

Table 6
Accenture's Results From the CheckThat! 2023 Lab Task 2

Language	Class	Precision	Recall	F1-score
Arabic	OBJ	0.936	0.810	0.869
	SUBJ	0.473	0.756	0.582
	macro avg	0.705	0.783	0.725
	weighted avg	0.851	0.800	0.816
English	OBJ	0.630	0.879	0.734
	SUBJ	0.827	0.528	0.644
	macro avg	0.728	0.703	0.689
	weighted avg	0.733	0.696	0.687
Turkish	OBJ	0.841	0.667	0.744
	SUBJ	0.757	0.892	0.819
	macro avg	0.799	0.779	0.781
	weighted avg	0.796	0.788	0.784
German	OBJ	1.000	0.005	0.010
	SUBJ	0.335	1.000	0.501
	macro avg	0.667	0.503	0.256
	weighted avg	0.778	0.337	0.174
Italian	OBJ	0.866	0.681	0.763
	SUBJ	0.446	0.709	0.548
	macro avg	0.656	0.695	0.655
	weighted avg	0.754	0.689	0.706
Dutch	OBJ	0.877	0.380	0.531
	SUBJ	0.578	0.941	0.716
	macro avg	0.728	0.661	0.623
	weighted avg	0.735	0.646	0.618

Table 7
Accenture's Results from the CheckThat! 2023 Lab Task 2

Language	Accuracy
Arabic	0.800
English	0.696
Turkish	0.788
German	0.337
Italian	0.689
Dutch	0.646

Table 8
BERT-based Arabic Model Performance at Different Level of Data Augmentation.

Augmentation	Class	Sample size	Precision	Recall	F1-score
No augmentation	OBJ	905	0.932	0.835	0.881
	SUBJ	280	0.500	0.732	0.594
	macro avg		0.716	0.783	0.737
	weighted avg		0.853	0.816	0.828
AR > EN > AR	OBJ	905	0.949	0.826	0.884
	SUBJ	560	0.512	0.805	0.626
	macro avg		0.731	0.816	0.755
	weighted avg		0.869	0.823	0.836
AR > EN > AR, and	OBJ	905	0.935	0.838	0.884
AR > EN > ES > EN > AR	SUBJ	840	0.508	0.744	0.604
	macro avg		0.722	0.791	0.744
	weighted avg		0.857	0.820	0.832
AR > EN > AR,	OBJ	905	0.936	0.810	0.869
AR > EN > ES > EN > AR, and	SUBJ	1,120	0.473	0.756	0.582
AR > EN > FR > EN > AR	macro avg		0.705	0.783	0.725
	weighted avg		0.851	0.800	0.816

Table 9
BERT-based Arabic Model Performance at Different Level of Data Augmentation.

Augmentation	Accuracy
No augmentation	0.816
AR > EN > AR	0.823
AR > EN > AR, and AR > EN > ES > EN > AR	0.820
AR > EN > AR, AR > EN > ES > EN > AR, and AR > EN > FR > EN > AR	0.800

6. Discussion

We observe that a specialized style-classification model outperformed the RoBERTa-large model quite significantly as seen in Table 10 and 11. This is likely because for a subjectivity classification task there is a heavy emphasis on vocabulary and terminology, which is a lacking in the relatively small training set provided. The raw RoBERTa did not have enough training vocabulary to outperform a specialized model. We also observe a diminishing return when over augment with the Arabic training set. As mentioned before, vocabulary plays a key role and augmenting with several pivot languages may have affected the data quality, potentially removing keywords that determine subjectivity. Look at the example below of a document labeled subjective after only one translation from Arabic to English:

"Are there **any resolutions** that the Security Council **may issue** to ensure that Egypt's water

Table 10
Performance Comparison Between RoBERTa and BERT-based Specialized Style Classification Model

Model	Class	Precision	Recall	F1-score
RoBERTa-large	OBJ	0.630	0.879	0.734
	SUBJ	0.827	0.528	0.644
	macro avg	0.728	0.703	0.689
	weighted avg	0.733	0.696	0.687
cffl/bert-base-styleclassification-subjective-neutral	OBJ	0.844	0.655	0.738
	SUBJ	0.739	0.890	0.807
	macro avg	0.792	0.773	0.773
	weighted avg	0.789	0.778	0.774

Table 11
Performance comparison between RoBERTa and BERT-based specialized style classication model

Model	Accuracy
RoBERTa-large	0.696
cffl/bert-base-styleclassification-subjective-neutral	0.778

share in the Nile River will not be affected?"

The second round of back-translation (Arabic > English > Spanish > English) then produces:

"Is there **a resolution** that the Security Council **can issue** to ensure that Egypt's water quota in the Nile River is not affected?"

And the third (Arabic > English > French > English) produces:

"Are there **resolutions** that the Security Council **could adopt** to ensure that Egypt's share of water in the Nile is not affected?"

By the second or third translation, the tone of the statement has shifted towards much more objective. This results in much lower model performance. We can see the results of these experiments in Table 8.

Due to extremely low sample size on the subjective class, we augmented Arabic and Italian training data three times. Table 12 shows the average cosine similarity score between each translation results to the original and the weighted average sentiment score of the pivoting English back-translation based on the Vader Lexicon [17]. For Arabic, there was no notable difference between the scores. However, for Italian, cosine similarity shows small decreases as more layers of back-translation are added, indicating a small level of semantic drift. Additionally, mean sentiment score decreases indicating subjectivity-level of the lexicon decreases as well.

Our paper suggests there may be a 'sweet spot' in BLEU score for data agumentation for

Table 12
Average Cosine Similarity of Italian Back-translation Compared to the Original and Weighted Average English Vader Sentiment Score

Language	Back-translation	Avg. Cosine Similarity	Avg. Sentiment Score
Arabic	AR > EN > AR	0.536	0.022
	AR > EN > ES > EN > AR	0.513	0.025
	AR > EN > FR > EN > AR	0.511	0.055
Italian	IT > EN > IT	0.562	0.097
	IT > EN > ES > EN > IT	0.492	0.074
	IT > EN > FR > EN > IT	0.466	0.032

back-translation, where a perfect translation would not add sufficient noise to the training data and a poor translation would not add sufficient context. We would recommend exploration of the BLEU score space as an optimization problem in future work.

7. Conclusion

We have described the back-translation augmentation strategies and models employed by Team Accenture's submissions to Task 2. Team Accenture's back-translation and foundation model approach yielded the 3rd highest submissions in Arabic, 4th in Turkish, 5th in Dutch, and 8th in German and English. In future work, we hope to explore in more detail to what extent back-translation data augmentation can inhibit subjectivity detection systems.

References

[1] E. Cambria, S. Poria, A. Gelbukh, M. Thelwall, Sentiment analysis is a big suitcase, IEEE Intelligent Systems 32 (2017) 74–80. doi:10.1109/MIS.2017.4531228.

[2] I. Chaturvedi, E. Cambria, R. E. Welsch, F. Herrera, Distinguishing between facts and opinions for sentiment analysis: Survey and challenges, Information Fusion 44 (2018) 65–77.

[3] L. L. Vieira, C. L. M. Jeronimo, C. E. Campelo, L. B. Marinho, Analysis of the subjectivity level in fake news fragments, in: Proceedings of the Brazilian Symposium on Multimedia and the Web, 2020, pp. 233–240.

[4] C. L. Jeronimo, L. B. Marinho, C. E. Carmpelo, A. Veloso, A. S. da Costa Melo, Characterization of fake news based on subjectivity lexicons., J. Data Intell. 1 (2020) 419–441.

[5] P. Kasnesis, L. Toumanidis, C. Z. Patrikakis, Combating fake news with transformers: A comparative analysis of stance detection and subjectivity analysis, Information 12 (2021) 409.

[6] A. Galassi, F. Ruggeri, A. B.-C. no, F. Alam, T. Caselli, M. Kutlu, J. M. Struss, F. Antici, M. Hasanain, J. Köhler, K. Korre, F. Leistra, A. Muti, M. Siegel, M. D. Turkmen, M. Wiegand, W. Zaghouani, Overview of the CLEF-2023 CheckThat! lab task 2 on subjectivity in news

articles, in: Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum, CLEF ’2023, Thessaloniki, Greece, 2023.

[7] E. Williams, P. Rodrigues, S. Tran, Accenture at CheckThat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation, 2021. arXiv:2107.05684.

[8] G. A. Aye, S. Kim, H. Li, Learning autocompletion from real-world datasets, in: 2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), IEEE, 2021, pp. 131–139.

[9] 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).

[10] 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, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[11] 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).

[12] S. Schweter, BERTurk - BERT models for turkish, 2020. URL: https://doi.org/10.5281/zenodo.3770924. doi:10.5281/zenodo.3770924.

[13] B. Chan, S. Schweter, T. Möller, German’s next language model, 2020. arXiv:2010.10906.

[14] S. Schweter, Italian BERT and ELECTRA models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[15] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, M. Nissim, Bertje: A Dutch BERT model, 2019. arXiv:1912.09582.

[16] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, 2017. arXiv:1703.01365.

[17] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, Proceedings of the International AAAI Conference on Web and Social Media 8 (2014) 216–225. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550. doi:10.1609/icwsm.v8i1.14550.