Performance of two French BERT models for French language on verbatim transcripts and online posts

Emmanuelle Kelodjoue and **Jérôme Goulian** and **Didier Schwab**
University of Grenoble Alpes, LIG
Bâtiment IMAG Université Grenoble Alpes
700, avenue centrale 38401 Saint Martin d'Hères
emmanuelle.kelodjoue-nguemegne@univ-grenoble-alpes.fr
jerome.goulian@univ-grenoble-alpes.fr
didier.schwab@univ-grenoble-alpes.fr

Abstract

Pre-trained models based on the Transformer architecture have achieved notable performances in various language processing tasks. This article presents a comparison of two pre-trained versions for French in a three-class classification task. The datasets used are of two types: a set of annotated verbatim transcripts from face-to-face interviews conducted during a market study and a set of online posts extracted from a community platform. Little work has been done in these two areas with transcribed oral corpora and online posts in French.

1 Introduction

Opinion mining has recently undergone a change with the rise of deep learning and, especially, the use pre-trained Language Models (Vaswani et al., 2017). The use of the latter such as ELMO (Peters et al., 2018) and BERT (Devlin et al., 2018) has led to significant improvements on a wide range of NLP tasks for the English language, from relation extraction to document classification (Peng et al., 2019; Laskar et al., 2020). French variants such as FlauBERT (Le et al., 2020) and CamemBERT (Martin et al., 2019) were proposed later on.

In this work, we are interested in the classification of two types of data as being either *in favour* (motivation), either *not in favour* (barriers) or *in favour on the condition that* (condition) :

- Verbatim transcripts from face-to-face interviews conducted in the context of a market potential study of an innovative product using natural language processing methods (NLP).
- Online posts comes from a community platform called *Yoomaneo*.[1]

Since we work with French data, we propose to compare and analyse the performance of two pre-trained versions for French. Additionally, since the collected data is small, we propose to augment the data with different augmentation techniques.

Contribution: This paper aims to compare and analyse the performances of two french BERT models on two different types of data.

2 Dataset : Constitution and Annotation

2.1 Dataset origin: Verbatim transcript

The dataset used to build and evaluate the French BERT models in this work comes from a set of 4367 verbatim. These verbatim were manually extracted from 75 transcripts[2] of face-to-face interviews. [3] To use this dataset for our research task, we conduct a human evaluation. We gather 6 evaluators and create two subunits of 3 annotators and add one more[4] to balance the evaluation of the two groups. We ask each group to review monthly 200 verbatim from the 4367.

Evaluation rules:

Only the verbatim whose classification obtained an interrater agreement according to the following rules were kept. Each verbatim of our initial corpus (4367) must be evaluated by at least 3 people. If a class (barrier, motivation or condition) results in an agreement greater than or equal to 50% for a verbatim and there is not a 50/50 on it, the selected verbatim and the assigned class is selected. On the other hand, if the interrater agreement is less than 50% or if there is 50/50 on two labels, the verbatim is eliminated from the corpus. 1578 out of 4367 verbatim transcripts have been evaluated, and only 839 verbatim transcripts obtained an agreement

[1] https://www.yoomaneo.com/

[2] 434 081 tokens.

[3] The interviews were conducted as part of different market potential studies catering various innovative products in the field of electricity, health, electrical goods, gerontology, automatism and pastry.

[4] We called him the *common annotator* since his role is to fill the empty space left by one of the six initials annotators.

greater than or equal to 50%. The distribution of the corpus is given in Table 1.

Classes	Number of verbatim
Barriers	189
Motivation	407
Condition	243

Table 1: Number of verbatim per categories.

2.2 Dataset origin: Online posts

Yoomaneo is a free community platform open to all. It was created in 2020 by the company Ixiade.[5] Yoomaneo was created to build a database of individuals willing to participate in studies on Innovation. Ixiade is responsible for the recruitment of the participants of the studies, who are then invited to download the application. For our case, 755 responses or posts were extracted from Yoomaneo. These posts come from 4 different projects which focus on the evaluation of different innovative concepts in 3 different domains: health, well-being and electrical (2 projects).

Evaluation rules The collected posts were then given to 3 research fellows to evaluate. The evaluation procedure is similar to the one mentioned in section 2.1. Only the posts which received at least the same evaluation (same category when annotated) were kept. As a result, of the **755 evaluated**, **433** were assigned to the *motivation* class, **112** to the *barrier* class, **97** to the *condition* class, **65** were deemed unclassifiable, and **48** received no agreement. The distribution of the corpus is given in Table 2.

Classes	Number of verbatim
Barriers	112
Motivation	433
Condition	97

Table 2: Number of verbatim per categories.

3 Data Augmentation

Data amplification involves all the techniques for amplifying the amount of data available by adding slightly modified copies of the original data (Li et al., 2021) or artificially generating data from the original data through transformations (Taylor and Nitschke, 2018) with the goal of increasing the size of the dataset. It has been used in various fields such image classification (He et al., 2016), speech recognition (Park et al., 2019), etc. In this work, 4 different popular augmentation methods have been implemented and adapted for text classification for the French language (Bayer et al., 2021): synthetic noise (Feng et al., 2020; Belinkov and Bisk, 2017), synonym replacement (Wei and Zou, 2019; Feng et al., 2020; Coulombe, 2020), random trio techniques (Feng et al., 2020) and back-translation (Mercadier, 2020; Marivate and Sefara, 2020), (Feng et al., 2020), (Wei and Zou, 2019), and (Marivate and Sefara, 2020). To our knowledge, most of the mentioned techniques have only been applied to English data reviews and not on the type of data this work used: verbatim transcripts and online posts.

3.1 Synthetic noise

For each verbatim transcript in our training dataset, we randomly delete, insert and swap characters according to a replacement percentage rate. We produce for a verbatim transcript 5%, 10%, and 15% noise variations.

3.2 Random trio techniques

For random trio techniques, we randomly remove a word which is not a stopword, insert a random synonym of a word into a random position in the verbatim transcript and swap the position of two words with a percentage rate of 5% (5% of the words are changed).

3.3 Replacement methods

Lexical replacement approach is a technique that replaces a word or words in a text with similar words. Most works (Kolomiyets et al., 2011; Zhang et al., 2015) replace words in the original text with their synonyms using WordNet (Esuli and Sebastiani, 2007). Since we deal with French data, we used the lexical resource DBnary (Sérasset, 2012; Sérasset and Tchechmedjiev, 2014). DBnary is a large lexical resource which provides multilingual lexical data extracted from Wiktionary. The dataset contains extracts from 22 Wiktionary languages. We replace only adjectives, adverbs, verbs and nouns with a randomly chosen synonym of the same POS provided by DBnary. We use Stanza (Qi et al., 2020) for tagging.

[5] https://www.ixiade.com/

3.4 Back-translation

Back-translation (Sennrich et al., 2015) consists in translating a sentence from a source language to a target language. The sentence obtained after translation from the source language to the target language is then translated back into the source language. This approach makes it possible to obtain different variants of the same sentence. We use Deepl[6] translation service web to produce those new data for our training dataset. We used all the languages provided by Deepl, approximately 25 languages.

Method	Text
Original	Tout à fait. Après il peut y avoir une application pour les IPAD, et une autre pour les smart phone, c'est pas le même usage.
Synthetic Noise	Tout à faeit. Après il **put** y avoir **upne applictaiown** pour **lhes IaPAD**, et une autre pour les smart phone, **cv'est** pas le même usage.
Random trio	Tout à fait . Après il peut y avoir une **usage** pour les IPAD, et une autre pour les smart phone , c' est pas le même **application**.
Synonym replacement	Tout à fait . Après il peut y avoir une application pour les IPAD , et une autre pour les smart phone , c' est pas le même **emploi** .
Back-translation	C'est vrai. S'il existe une application pour iPad et une application pour smartphone, il ne s'agit pas du même travail.

Table 3: Example of a verbatim transcript and its variations using our augmentation methods. Changes are bolded.

4 Experimental Setup

We chose 4 data augmentation techniques and 2 Pretrained Models (FlauBERT and CamemBERT) for this experimental work.

4.1 Data splitting and augmentation

Methods	Training	Testing
Original	503	168
Synthetic Noise	1981	168
Random trio	8024	168
Synonym replacement	6822	168
Back-translation	11 236	168

Table 4: Overview of the augmented datasets for the verbatim dataset.

We divide our dataset into 3 subsets: train, dev and test (respectively 60%, 20%, 20%). We augment only the training set. Table 3 gives an example of verbatim transcripts generated using the different augmentation methods mentioned above. Table 4 and 5 gives an overview of the training size per augmentation method.

Methods	Training	Testing
Original	384	129
Synthetic Noise	1465	129
Random trio	5481	129
Synonym replacement	3854	129
Back-translation	7691	129

Table 5: Overview of the augmented datasets for the online posts dataset.

4.2 Pretrained Models and Finetuning

Model description. FlauBERT (Le et al., 2020) is a French BERT model. It was trained on 71 GB of French text corpus. The corpus consists of 24 subcorpora covering diverse topics and writing styles from formal and well-written text (e.g. Wikipedia and books).[7] CamemBERT is also a language model for French based on the RoBERTa (Liu et al., 2019) architecture pretrained on the French corpus OSCAR (Suárez et al., 2019) (138 GB) and CCNET (Wenzek et al., 2019) (135 GB). Both FlauBERT and CamemBERT were trained on the masked Language Modeling (MLM) task.

Architecture. For our task, we append the relevant predictive layer on top of CamemBERT's and FlauBERT's architecture. We fine-tune all the different models to follow the process described by Devlin et al. (2018) and followed by Le et al. (2020). The classification head for FlauBERT consists of the following layers, a dropout, a linear layer followed by the activation function tanh, a dropout and another linear layer. To obtain the probabilities for each class, the softmax function was used. The dimensions of the inputs of the linear layers are respectively equal to the size of the Transformer. For CamemBERT, the classification heads are the same as the ones described in Martin et al. (2019).

Parameters. As far as the hyperparameters are concerned, they are all fixed at the time of learning, with a batch size of 8 for all the architectures. The number of epochs is set to 5 and the learning rate to 5e-5 for the first epoch, then decreasing linearly. The AdamW (Kingma and Ba, 2014) optimizer is used.

[6] https://www.deepl.com/fr/translator

[7] http://www.gutenberg.org.

5 Results and Analysis

In this section, we present the results on our two test data. We compare the performance of FlauBERT with its competitor (CamemBERT). The metrics used to measure the performance of each method were the F1-score and the accuracy (F-micro). The F-score is used as metric since our data are imbalanced in order to observe the real performance of the model. The results are evaluated according to the amplification method used and the architecture used. Our baseline is the model without amplification.

5.1 FlauBERT

For FlauBERT, we use the 3 model sizes: FlauBERT BASE CASED (BC), FlauBERT BASE UNCASED (BU) and FlauBERT LARGE (L). Table 6 presents the size of data on which each model was trained.

Model	Parameters	Architecture	Training corpus
FlauBERT BASE CASED (BC)	138M	Base	24 corpora (71GB)
FlauBERT BASE UNCASED (BU)	137M	Base	24 corpora (71GB)
FlauBERT LARGE (L)	373M	Large	24 corpora (71GB)

Table 6: pre-trained model size for FlauBERT (Le et al., 2020).

TAD	Verbatim transcripts					
	FlauBERT Base Cased		FlauBERT Base Uncased		FlauBERT Large	
	accuracy	F1	accuracy	F1	accuracy	F1
0 - Baseline	0.482	0.217	0.500	0.267	0.589	0.538
1 - BT	0.667	0.604	0.690	0.650	**0.690**	**0.657**
	+0.18	+0.39	+0.19	+0.38	**+0.10**	**+0.12**
2 - SR	0.589	0.574	0.649	0.607	0.690	0.641
	+0.11	+0.36	+0.15	+ 0.34	+0.10	+0.10
3 - RT	0.595	0.558	**0.714**	**0.683**	0.583	0.411
	+0.11	+0.34	**+0.21**	**+0.42**	-0.01	-0.13
4 - NI	**0.673**	**0.591**	0.685	0.641	0.625	0.514
	+0.19	**+0.37**	+0.18	+0.37	+0.04	-0.02

Table 7: FlauBERT: F1 and accuracy score for verbatim transcripts test data.

Table 7 presents the final accuracy and F1 on test set for the verbatim transcripts. The results show that FlauBERT BU performs better than the CASED model and the LARGE model, with an accuracy score of 0.714 and F1 score of 0.682. Overall, Back-translation and noise injection perform better for all the 3 models, with an average accuracy greater than 0.60. Huge improvement is observed with the F1 score for all the models, except for the case where FlauBERT LARGE is combined with random trio and Noise Injection. One reason may be that too much injection and replacement of words might have altered the semantic sense of the training data when augmenting it.

TAD	Online Posts					
	FlauBERT Base Cased		FlauBERT Base Uncased		FlauBERT Large	
	accuracy	F1	accuracy	F1	accuracy	F1
0 - Baseline	0,667	0,269	0,674	0,269	0,667	0,267
1 - BT	0,698	0,514	0,791	0,660	**0,822**	**0,750**
	+0,03	+0,24	+0,12	+0,39	**+0,15**	**+0,48**
2 - SR	0,713	0,582	0,752	0,621	**0,829**	**0,733**
	+0,05	+0,31	+0,08	+0,35	**+0,16**	**+0,47**
3 -RT	**0,736**	**0,648**	0,721	0,515	0,814	0,723
	+0,07	**+0,38**	+0,05	+0,25	+0,15	+0,46
4 - NI	0,651	0,484	**0,798**	**0,714**	0,829	0,729
	-0,02	+0,22	+0,12	+0,44	+0,16	+0,46

Table 8: FlauBERT: F1 and accuracy score for online posts test data.

Table 8 presents the results on the test set for online posts. The results show that FlauBERT L performs slightly better than the CASED model and the LARGE model, with an accuracy score greater than 0.80 for all the amplification methods. The best score is obtained with synonym replacement and FlauBERT L with an accuracy score of 0.829 and F1 of 0.733.

By comparing the results, we observe that the amplification methods combined with the different FLauBERT models improve the classification task for both test data. Nevertheless, the results are more significant on the online post data, with an accuracy above 0.80. This might be because they are somewhat similar to reviews or critics. Verbatim transcripts are quite particular since they come from oral dialogue which has been transcribed and revised. The classification models have somewhat more difficulties to classify those type of data compare to online posts, even though the accuracy is quite good (> 0.70 on verbatim transcripts test data). Random trio and synonym replacement are respectively the ones which produced the best score for the test data for verbatim transcripts and test data for online posts.

In the next section, we present the results obtained when using CamemBERT model.

5.2 CamemBERT

For CamemBERT, we used three model sizes which were introduced in Martin et al. (2019): CamemBERT BASE O for the model trained on the OSCAR corpus, CamemBERT BASE C for the model trained on the CCNET corpus and CamemBERT LARGE trained of the CCNET corpus.

Table 10 presents the final accuracy and F1 on test set for online posts. The results show that CamemBERT LARGE performs better than the BASE model, with an accuracy score of 0.756 and

Model	Parameter	Architecture	Training corpus
CamemBERT BASE O	110M	Base	corpus OSCAR (135 GB)
CamemBERT LARGE	335M	Large	corpus CCNet (135 GB)
CamemBERT BASE C	110M	Base	corpus CCNet (135 GB)

Table 9: Pre-trained models size for CamemBERT (Martin et al., 2020).

TAD	Verbatim transcripts					
	CamemBERT B (OSCAR)		CamemBERT B (CCNet)		CamemBERT L (CCNet)	
	accuracy	F1	accuracy	F1	accuracy	F1
0 - Baseline	0,482	0,217	0,607	0,458	0,494	0,318
1 - BT	0,696 +0,21	0,640 +0,42	**0,732** **+0,13**	**0,687** **+0,47**	0,673 +0,18	0,611 +0,39
2 - SR	**0,714** **+0,23**	**0,663** **+0,45**	0,702 +0,10	0,653 +0,44	0,649 +0,15	0,581 +0,36
3 - RT	0,714 +0,23	0,648 +0,43	0,655 +0,05	0,594 +0,38	**0,756** **+0,26**	**0,720** **+0,50**
4 - NI	0,685 +0,20	0,642 +0,43	0,667 +0,06	0,621 +0,16	0,714 +0,22	0,687 +0,37

Table 10: CamemBERT: F1 and accuracy score for Verbatim transcripts test data.

F1 score of 0.720. Random trio is the best performing method (acc.: 0.756) follow by the back-translation method (acc.: 0.732) and synonym replacement (acc.: 0.714).

TAD	Online Posts					
	CamemBERT B (OSCAR)		CamemBERT B (CCNet)		CamemBERT L (CCNet)	
	accuracy	F1	accuracy	F1	accuracy	F1
0 - Baseline	0,674	0,269	0,752	0,484	0,674	0,269
1 - BT	**0,783** **+0,11**	**0,68** **+0,41**	0,814 +0,06	0,755 +0,27	0,822 +0,15	0,771 +0,50
2 - SR	0,767 +0,09	0,672 +0,40	0,845 +0,09	0,759 +0,27	**0,868** **+0,19**	**0,801** **+0,53**
7 - RT	0,744 +0,07	0,666 +0,40	**0,853** **+0,10**	**0,780** **+0,30**	0,775 +0,10	0,708 +0,44
8 - NI	0,744 +0,07	0,580 +0,31	0,806 +0,05	0,731 +0,25	0,806 +0,13	0,624 +0,36

Table 11: CamemBERT: F1 and accuracy score for Online Posts test data.

Table 11 show that CamemBERT LARGE performs better than the BASE model, with an accuracy score of 0.868 and F1 score of 0.801. Random trio is the best performing method, follow by the back-translation method (acc.: 0.853) and synonym replacement (acc.: 0.783).

By comparing the results, we observe that augmentation methods used in this work clearly improved the performances for both CamemBERT and FlauBERT. Overall, CamemBERT performances are better than FlauBERT. Synonym replacement combined with CamemBERT LARGE is the best performing duo on verbatim and online posts test data. We also noted that performances on post online are better than on verbatim transcripts. One reason may be the type of data the pre-trained model were trained on. CCNET corpus were crawled from internet, so they may be more similar or linguistically closer to online posts than verbatim transcripts. A linguistic analysis of the data used to trained CamemBERT model may be interesting to conduct in order to explore the lingusitic similarities or differences with our datasets. In conclusion, the results are promising and clearly open up work prospects.

6 Conclusion

We have presented a work where we sought to compare the performances of two BERT models for French language on a three-class classification task . Firstly, we show that simple augmentation techniques used for text classification can be implemented and adapted for the datasets used in this work. Overall, we also obseved that CamemBERT model was better than FlauBERT for this task and the best amplification method was synonym replacement. For future works, we would like to use other pretrained language models for French such as XLNET, BERT multilingual, etc. In this paper, we just focus on comparing two French Variants. We also think exploring the linguistic features of our dataset in the training of the model may be interesting with the goal of evaluating their impact on the performance. Finally, we also think that trying to other amplification methods such as replacement via a language model may be interesting.

The data used in this work comes from a private enterprise, and we have not received their consent to share the dataset.

References

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2021. A survey on data augmentation for text classification. *arXiv preprint arXiv:2107.03158*.

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Claude Coulombe. 2020. *Techniques d'amplification des données textuelles pour l'apprentissage profond*. Ph.D. thesis, Télé-université.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Andrea Esuli and Fabrizio Sebastiani. 2007. Sentiwordnet: a high-coverage lexical resource for opinion mining. *Evaluation*, 17:1–26.

Steven Y Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. 2020. Genaug: Data augmentation for finetuning text generators. *arXiv preprint arXiv:2010.01794*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2011. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, volume 2, pages 271–276. ACL; East Stroudsburg, PA. En ligne à l'adresse : https://scholar.google.com/scholar?hl=fr&as_sdt=0%2C5&q=kolomiyets+synonym&btnG=.

Md Tahmid Rahman Laskar, Xiangji Huang, and Enamul Hoque. 2020. Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5505–5514.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Bohan Li, Yutai Hou, and Wanxiang Che. 2021. Data augmentation approaches in natural language processing: A survey. *arXiv preprint arXiv:2110.01852*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Vukosi Marivate and Tshephisho Sefara. 2020. Improving short text classification through global augmentation methods. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 385–399. Springer.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Yves Mercadier. 2020. *Classification automatique de textes par réseaux de neurones profonds: application au domaine de la santé*. Ph.D. thesis, Université Montpellier.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Gilles Sérasset. 2012. Dbnary: Wiktionary as a lmf based multilingual rdf network. In *Language Resources and Evaluation Conference, LREC 2012*.

Gilles Sérasset and Andon Tchechmedjiev. 2014. Dbnary: Wiktionary as linked data for 12 language editions with enhanced translation relations. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, pages 67–71.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Luke Taylor and Geoff Nitschke. 2018. Improving deep learning with generic data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1542–1547. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657. En ligne à l'adresse : https://proceedings.neurips.cc/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.