BERTweetFR : Domain Adaptation of Pre-Trained Language Models for French Tweets

Yanzhu Guo
École Polytechnique, France
Shanghai Jiao Tong University, China
yanzhu.guo@polytechnique.edu

Virgile Rennard
École Polytechnique, France
virgile@rennard.org

Christos Xypolopoulos
École Polytechnique, France
christos.xypolopoulos@polytechnique.edu

Michalis Vazirgiannis
École Polytechnique, France
mvazirg@lix.polytechnique.fr

Abstract

We introduce **BERTweetFR**, the first large-scale pre-trained language model for French tweets. Our model is initialized using the general-domain French language model CamemBERT (Martin et al., 2020) which follows the base architecture of BERT. Experiments show that BERTweetFR outperforms all previous general-domain French language models on two downstream Twitter NLP tasks of offensiveness identification and named entity recognition. The dataset used in the offensiveness detection task is first created and annotated by our team, filling in the gap of such analytic datasets in French. We make our model publicly available in the transformers library with the aim of promoting future research in analytic tasks for French tweets.

1 Introduction

Vector representations of words have given rise to the application of deep learning methods in NLP. Traditional pre-training approaches, such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) are static, learning a single representation for every word regardless of context. However, words are often polysemous with different meanings depending on the context. More recently, models are trained to integrate contextual meaning : the output word embeddings depend on the whole input sequence rather than only the word itself. While the idea was initially implemented with recurrent neural networks (Dai and Le, 2015) (Ramachandran et al., 2017), recent models have predominantly been based on the transformers architecture (Vaswani et al., 2017), with BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) among the most popular. These contextualized word representation models have opened new doors for researchers as they can be applied to numerous downstream tasks by fine-tuning or prompting. In fact, large-scale pre-trained language models have become the go-to tool for building new NLP applications, saving researchers from the enormous amount of computational resource and data required for training model weights from scratch.

In the past few years, human society has become more digitally connected than ever before. People use social media to report the latest news, but also to express their opinions and feelings about real-world events. As one of the most popular micro-blogging platforms, Twitter has become a primary source for social media user-generated data (Ghani et al., 2019). However, tweets are more often written in informal language compared to the carefully edited texts that are published in traditional data sources such as Wikipedia and printed media. They have their own set of features such as the recurrent use of irregular or abbreviated words, the large quantity of spelling or grammatical mistakes, the employment of improper sentence structures and the occurrence of mixed languages (Farzindar and Inkpen, 2020). This presents challenges to standard NLP methods when applied to Twitter data.

Domain-adaptive pre-training has been revealed to provide significant gains in helping models encode the complexity of specific textual domains (Gururangan et al., 2020). While efforts on domain adaptation of large-scale language models to Twitter language have been made in English (Nguyen et al., 2020), there has been no similar work in any other language.

We start addressing this problem by releasing

BERTweetFR, a pre-trained language model for French tweets together with a manually labeled dataset for offensiveness identification in French tweets. We evaluate our model on two downstream tasks : offensiveness identification and named entity recognition.

We compare the performance of our model to the best-performing transformer-based models pre-trained on general domain French texts, namely CamemBERT (Martin et al., 2020), FlauBERT (Le et al., 2020) and BARThez (Eddine et al., 2020). Experiments show that our model outperforms all three of them on both of the downstream tasks.

In order to facilitate future research on French tweets, we make our model publicly available on Huggingface's model hub [1] as well as on our team's website dedicated to French linguistic resources [2]. We will also release our offensiveness identification dataset while respecting the limitations of Twitter's developper policy.

2 Building the Model

In this section, we describe the collection steps and pre-procesing pipeline for our pre-training dataset, present the model architecture and introduce the training objective and optimization setup.

2.1 Pre-training Data

We use a 16GB dataset of 226M deduplicated French tweets. The tweets are deduplicated using open-source tool runiq[3]. In addition, we filter out tweets with fewer than 5 tokens assuming they do not contain substantial information. The average length of a tweet is 30 tokens.

2.1.1 Data Collection

Our final dataset for pre-training is an aggregation of three corpora from different sources. The aggregation of these corpora makes this dataset the largest one for French tweets up to this date. It is also beneficial to aggregate different corpora in order to cover tweets from different time periods with diverse topics and styles.

The three sources we use are as follows :

- We start by downloading tweets from the general Twitter Stream[4] grabbed by the Archive Team, containing of tweets streamed from January 2016 to December 2019. Selecting only the French tweets with Twitter's built-in feature, we obtain a corpus of $34M$ unique tweets.

- We also build a COVID-19 related corpus of French tweets relevant to the COVID-19 pandemic posted between September 2020 and April 2021. In this case, our filters are focused on tweets that include the hashtags "covid19" and "coronavirus". Through Twitter's public streaming API, we extracted tweets in French marked with either or both of the two above hashtags. This corpora consists of $19M$ unique tweets.

- Finally, we make use of a previous Twitter dataset constructed for socioeconomic analysis (Abitbol et al., 2018). This corpus includes a collection of tweets in French between the years 2014 and 2018. We extract $173M$ unique tweets form this corpus.

2.1.2 Data Pre-Processing

We only implement minimal data cleaning before inputting the sequences into the tokenizer. Following (Nguyen et al., 2020), we normalize the Tweets by converting user mentions and web/url links into special tokens @USER and HTTPURL.

For tokenization, we apply the CamemBERT tokenizer (Martin et al., 2020). The CamemBERT tokenizer segments input sequences into subword units using the SentencePiece (Kudo and Richardson, 2018) algorithm. This algorithm is an extension of Byte-Pair encoding (BPE) (Shibata et al., 1999) and WordPiece (Kudo, 2018) that eliminates the pre-tokenization step, thus more generally applicable. The vocabulary size is $32k$ subword tokens.

2.2 Model Architecture

BERTweetFR is initialized from the base version of CamemBERT. We choose to further fine-tune this model instead of starting from scratch because domain-adaptive pre-training have been proven to give very satisfying results in numerous downstream tasks spanning across a wide range of domains (Gururangan et al., 2020). The choice to

[1] https://huggingface.co/Yanzhu/bertweetfr-base

[2] http://master2-bigdata.polytechnique.fr/FrenchLinguisticResources/bertweetFr

[3] https://github.com/whitfin/runiq

[4] https://archive.org/details/twitterstream

employ CamemBERT instead of other French language models is based on its overall best performance on downstream tasks experimented in previous works (Eddine et al., 2020).

CamemBERT applies the multi-layer bidirectional Transformer architecture. The base version uses the same architectures as the base version of BERT with 12 layers, 768 hidden dimensions and 12 attention heads, adding up to a total of $110M$ parameters. CamemBERT follows the same optimized pre-training approach as RoBERTa, the only difference is that it uses whole-word masking and the SentencePiece tokenization instead of WordPiece.

2.3 Training Objective

Our model is trained on the Masked Language Modeling (MLM) task. With any given input sequence, 15% of the tokens are chosen for possible replacement. Among the selected tokens, 80% are further selected to be replaced by the special <MASK> token, 10% remain unchanged and 10% are replaced by random tokens. Finally, the model is trained to predict the tokens replaced by <MASK> using cross-entropy loss. Following RoBERTa, we do not fix the whole set of masked tokens during pre-processing but select them dynamically during the training process. The data is thus augmented when training for multiple epochs.

2.4 Optimization Setup

We employ the CamemBERT implementation in the transformers library (Wolf et al., 2020). The maximum sequence length is set to be 128, generating approximately $226M \times 30/128 \approx 53M$ sequence blocks. Following (Gururangan et al., 2020), we optimize the model using Adam (Kingma and Ba, 2015) with a batch size of 1280 across 8 V100 GPUs (32GB each) and a peak learning rate of 0.0001. We pre-train the model for 20 epochs in about 8 days with a total of $53M \times 20/1280 \approx 83K$ training steps.

3 Downstream Task Datasets

We evaluate the performance of BERTweetFR on two downstream Twitter NLP tasks. The datasets used in these downstream tasks are either constructed by ourselves or obtained from shared tasks in past conferences.

3.1 Offensive Language Identification

Along with the outbreak of COVID-19 came severe disruption in the French society. Effects include the public holding the government accountable for certain ways in which the pandemic was handled, as well as a rise of hateful sentiment towards the Asian community. In fact, ever since the explosion of COVID-19 on a global scale, unrest has led to an increase of violent incidents towards people of Asian descent.

In response to this phenomenon, we have created a human annotated dataset for general offensiveness detection in Tweets collected during the COVID-19 pandemic. Our dataset contains 5786 French tweets among which 1301 have been labeled as offensive. Offensiveness is not straightforward and can be subjective. In our labeling procedure, we consider a tweet as offensive in cases where personal attacks are detected. For example, "The chinese virus is tiring" would not be considered offensive, while "I hate the chinese for bringing us the chinese virus" would be. This is a binary sequence classification task. We randomly sample a 70/15/15 training/validation/test split with each class proportionally represented in each part of the split.

3.2 Named Entity Recognition

For the NER task, we take data from the CAp 2017 challenge (Lopez et al., 2017). This challenge proposes a new benchmark for the problem of NER for tweets written in French. The tweets were collected using the publicly available Twitter API and annotated with 13 types of entities : person, musicArtist, organisation, geoLoc, product, transportLine, media, sportsTeam, event, tvShow, movie, facility and other. Overall, the dataset comprises 6685 annotated tweets split into two parts: a training set consisting of 3000 tweets and a test set with 3685 tweets. For compatibility with previous research, the data were released tokenized using the CoNLL format and the BIO encoding.

4 Baselines

As our BERTweetFR model is the first pre-trained language model for French tweets, we compare it with the following general-domain language models for French.

CamemBERT As mentioned in 2.2, CaemBERT (Martin et al., 2020) is the model from which we

Table 1: Offensive Identification scores on the best model selected.

	CamemBERT	FlauBERT	BARThez	**BERTweetFR**
Accuracy	86.47	86.87	84.35	**88.07**
F1 Score	68.89	65.20	67.51	**71.27**

Table 2: NER score on the best model selected.

	CamemBERT	FlauBERT	**BERTweetFR**
Accuracy	94.78	94.73	**94.99**
F1 Score	61.01	60.57	**62.77**

started our fine-tuning for domain adaptation. It therefore serves as a natural baseline. Its architecture is already introduced in 2.2.

FlauBERT FlauBERT (Le et al., 2020) is another transfomer-based model trained on a very large and heterogeneous French corpus. It basically follows the same architecture as CamemBERT and is shown to outperform CamemBERT on some of the downstream tasks.

BARThez BARThez (Eddine et al., 2020) is the first French sequence-to-sequence model based on the base version of the BART architecture (Lewis et al., 2020). It has 6 encoder and 6 decoder layers with 768 hidden dimensions and 12 attention heads in both the encoder and the decoder. It is shown to be competitive in comparison with CamemBERT and FlauBERT.

5 Experiments and Results

In this section, we first describe our fine-tuning approaches. The baseline models follow the same fine-tuning procedures as BERTweetFR. We then report our results and compare with the baselines. As a result, our model substantially outperforms all baselines.

5.1 Offensive Language Identification

The offensive language identification task is a supervised sequence classification task. Following (Devlin et al., 2019), we append a linear prediction layer on top of the pooled output.

For fine-tuning, we employ transformers library to train BERTweetFR on the training set for 15 epochs. We use AdamW (Loshchilov and Hutter, 2019) with a fixed learning rate of 2.e-5 and a batch size of 32 following (Liu et al., 2019). We compute the classification accuracy and F1 score after each training epoch on the validation set, applying early stopping if their is no improvement after 3 consecutive epochs. We eventually select the model checkpoint with the highest F1 score to predict the final labels on the test set. Our results are listed in Table 1.

5.2 Named Entity Recognition

The NER task is a supervised token classification task. Following (Devlin et al., 2019), we append a linear prediction layer on top of the last Transformer layer with regards to the first subword of each word token.

For fine-tuning, we again employ transformers library to train for 30 epochs. We use AdamW (Loshchilov and Hutter, 2019) with a fixed learning rate of 2.e-5 and a batch size of 32, adding in weight decay. We compute performance scores for each entity class as well the overall F1 Micro score. We eventually select the model checkpoint with the highest F1 Micro score to predict the final labels on the test set. Our results are listed in Table 2. We do not compare with BARThez in this task because the original model is not implemented for sequence classification tasks.

6 Conclusion

In this work, we investigated the effectiveness of applying domain adaptation to the Twitter domain for large-scale pre-trained French language models. We demonstrate the value of our model showing that it outperforms all previous general-domain French language models on two downstream Twitter NLP tasks of offensiveness identification and named entity recognition.

Our contributions are as follows :

- We train and release the first large-scale pre-trained language model for French tweets : BERTweetFR. We make it publicly available in the transformers library and hope that it can facilitate and promote future research in analytic tasks for French tweets.

- We create and annotate the first dataset for offensiveness identification in French tweets. Such datasets already exist in several other languages and our effort fills in the gap for the French language.

- We create a framework and baseline for evaluating language models for French tweets. Datasets for Twitter tasks in French is very scarce and no previous work has ever combined different analytic tasks together in a unified framework.

With around 70% of all Twitter posts being in non-English languages, the lack of corresponding language models strongly hinders the community from exploiting the information contained in these valuable resources. For future work, we plan to train and release a series of such pre-trained language models for tweets in other low resource languages. We also call upon researchers from all over the world involved in natural language processing for social media to adapt language models in their respective languages and make them publicly available.

Acknowledgements

This research is supported by the French National research agency (ANR) via the ANR XCOVIF (AAP RA-COVID-19 V6) project. We would also like to thank the National Center for Scientific Research (CNRS) for giving us access to their Jean Zay supercomputer.

References

Jacob Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, and Eric Fleury. 2018. Socioeconomic dependencies of linguistic patterns in twitter: A multivariate analysis. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1125–1134, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Moussa Kamal Eddine, Antoine J-P Tixier, and Michalis Vazirgiannis. 2020. Barthez: a skilled pretrained french sequence-to-sequence model. *arXiv preprint arXiv:2010.12321*.

Anna Atefeh Farzindar and Diana Inkpen. 2020. Natural language processing for social media. *Synthesis Lectures on Human Language Technologies*, 13(2):1–219.

Norjihan Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem, and Ejaz Ahmed. 2019. Social media big data analytics: A survey. *Computers in Human Behavior*, 101:417–428.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit 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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer

Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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*.

Cédric Lopez, Ioannis Partalas, Georgios Balikas, Nadia Derbas, Amélie Martin, Coralie Reutenauer, Frédérique Segond, and Massih-Reza Amini. 2017. Cap 2017 challenge: Twitter named entity recognition. *arXiv preprint arXiv:1707.07568*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric 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*, pages 7203–7219, Online. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Prajit Ramachandran, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Copenhagen, Denmark. Association for Computational Linguistics.

Yusuke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.