Parlementarian Sentiment Analysis

Machine Learning for Natural Language Processing 2022

Yasmine Houri[*,1] and Mathis Sansu[†,1]

[1]ENSAE Paris, France

24 April 2022

Abstract

We perform a sentiment analysis on tweets of French MPs around the occurrence of exogenous shocks (first COVID-19 lockdown, Benalla scandal, French victory in the 2018 Football World Cup). Our main model is based on CamemBERT and yields a performance almost as good as human evaluation. This allows us to point out sociodemographic and political factors having an influence on the style of communication for MPs, as well as the fact that events can modify temporarily this communication framework.

Introduction. A traditional object of study in political sciences is the intertwining between the political and media spheres. With the development of digital social networks, new mechanisms in political communication seem to have emerged. Indeed, some talk about a "numerization" of the public space and view these social networks as new tools of communication for politicians. These new channels, according to some academic works, may enhance a personalization of political communication as well as a media disintermediation. We test in this paper the individual responses of politicians – French Members of Parliament (MPs) – to exogenous shocks in terms of their mode of communication: are they trying to convey positive or negative sentiments in reaction to chosen events, and are these responses correlated with sociodemographic or / and political characteristics of the individuals?

Data. We will work on the tweets of French MPs of the 15^{th} legislature of the V^{th} Republic (2017-2022). While they do not embody the political sphere as a whole, they constitute an important part of the national and local politics in France. Moreover, they constitute a population whose Twitter data are fairly accessible thanks to their official position and relatively small number and which is at the same time large enough to observe variations. We create a list of the Twitter handles of MPs thanks to several data sources[1]. We use the Academic Research Product Track V2 API Endpoint of Twitter[2] to retrieve the Twitter data. As a preliminary step, we make use of the R-package 'academictwitteR' (Barrie and Ho (2021)) to extract all tweets – this comprises tweets, retweets, quotes and replies - of MPs of the current legislature (2017-2022), using the function *get_all_tweets*[3]. After merging the database for all MPs, we obtain a complete database, where each line corresponds to a single tweet. From another academic database, we also joined sociodemographic and political variables about MPs. We dispose of variables informing at different levels, tweet-level or MP-level. To test our responses to shocks, we build three samples according to calendar window around events: one from mid-April to August 2018 centered around the Benalla scandal, one from June to July 2018 centered around the victory of France in the Football World Cup, one from March to May 2020 which encompasses the first COVID-19 lockdown.

Methods. The task we conduct in this work is one of sequence classification: it consists in sentiment analysis at the tweet-level, with 3 modalities: positive, negative, other. Since we work with corpora of tweets that are fairly recent, we will face Out-Of-Vocabulary (OOV) issues if we simply use *Word2vec* techniques. To avoid this, we make use of continuous word representations which are pre-trained on large unlabeled corpora. Thus, we will implement two models, a baseline model which relies on fastText (Bojanowski et al. (2017)) and a second model based on CamemBERT, a French adaptation of BERT (Martin et al. (2019); Devlin et al. (2018)). We focus on the discourses directly produced by MPs: our sample is thus restricted to tweets, quotes and replies in French[4] - according to Twitter's language classification.

In this study, we manually annotated 1,000 random tweets drawn form the initial database, each annotator doing so independently from one another. We then compared the agreement rate between annotators: we find a correspondence of 81.6%, which is around the standard for human annotation. This human annotation allows us to train our models towards the specific task we want to achieve - with the pre-trained embeddings - and to test the validity and robustness of our models. Before applying the models on our corpora, we implement a preprocessing function to clean undesirable features of tweets[5]. We use a classic machine learning approach to assess which model we will choose: we split our annotated sample into subsamples - train, test and / or validation sets - and we evaluate the performance of our model with the confusion matrix and *sklearn.metrics*' *classification_report* (precision, recall, f-score and accuracy).

The baseline model, relying on *fastText*, is based on subword embedding *i.e.* the vector representation of a word is an average of vector representations of character n-gram of the word. It allows to preserve some contextual information around the word, and also provides a vector representation even if the given word is OOV. The model was previously trained on Wikipedia and Crawl. To train our baseline model, we set the batchsize to 64. The input layer of our neural network is of size 200 according to the maximum length set during the padding step, the hidden dimension is of size 300, equal to the dimension of the pretrained vector, and the output dimension is of size 3, equal to the number of classes.

[*]yasmine.houri@ensae.fr

[†]mathis.sansu@ensae.fr

[1]database from previous academic research, scraping from websites https://www.nosdeputes.fr/ and https://www.assemblee-nationale.fr/, and manual updates.

[2]https://developer.twitter.com/en/portal/dashboard

[3]We can provide the R code on demand. The last update of this extraction was done the 21st of March.

[4]We restrict to French in order to avoid issues arising with multilingual corpus.

[5]We mainly remove emojis, symbols "@" and "#", as well as url links. We believe that they create more noise than they provide information for the sequence classification task.

The task-specific model is adapted from Bidirectional Encoder Representations from Transformers (BERT), which is a transformer-based language modelling and next-sentence prediction algortihm, developed by Google in 2018. BERT is the best performing algortihm for many NLP tasks at the moment. Its method incorporates Masked Language Modeling: BERT randomly masks some words in the given sentence, and uses bidirectional context prediction, *i.e.* it simultaneously uses previous and following words to predict the outcome. For our specific task, we use CamemBERT, a French version of BERT pre-trained on a non-shuffled version of the French sub-corpus of OSCAR[6], which amounts to 138Go of raw text. In order to use CamemBERT for downstream tasks, we fine-tune it, and extract pre-trained embedding vectors from it. We use the Adam optimizer since this one yields the better performance in terms of evaluation metrics.
BERT uses the Transformer encoder-decoder architecture. The basic version of BERT is made of 12 layers, each divided into 4 sublayers: attention, normalisation, feedforward, and normalisation. The two main pillars of Transformer time and space complexity are attention complexity and the memory needed for the number of layers. Let X the input of a recurrent attention layer, and (n, d) the shape of X, with n the number of word-vectors of dimension d. The complexity of a recurrent self-attention layer is equal to $O(n^2d + nd^2)$ (Vaswani et al. (2017)). However, we know that BERT is based on Multi-Head Attention layers encoder. What's more, CamemBERT is an adaptation of the post-BERT model RoBERTa, which has a longer pre-training period than BERT (Gillioz et al. (2020)). Therefore, our model is fairly more complex than BERT. Regarding space complexity, the more layers the NN has, the higher its space complexity will be, because the activations need to be stored in order to be used in the backpropagation.

Results. The baseline model is trained on the random manually annotated sample of size 1,000, using the Adam optimizer over 80 epochs, with a learning rate of 0.001. The overall accuracy on the test set is close to 0.50. Given that we have three classes, this model performs better than random.
The task-specific model is trained in the same way, but over 3 epochs instead, because it is deeper than the baseline model. The overall accuracy on the test set is equal to 0.77, which is very close to the accuracy of the human performance on this task (remember that we found a correspondence of 81.6% between our two independent manual annotations). In addition to accuracy, the prediction and recall of our model are also close to 0.70. We choose to go further with this model to perform the sentiment analysis. From this model, we predict the sentiment conveyed by each tweet for the three studied subsamples presented above.
Finally, a last step regarding our analyses consists into the interpretation of our results. This does not directly correspond to a Natural Language Processing task, but rather that the NLP tasks conducted before enable the production of data for these analyses[7]. We just present a few results and a few lines of inquiry for further research[8].
The application of the task-specific model sheds light on several relevant peaks of sentiment expression. Around the beginning of the first lockdown (Figure 1), the ratio of positive tweets reached 0.55, against 0.15 for negative tweets. Around the end of the first lockdown, the ratio of positive tweets rose to 0.61. The peak of the pandemic at the end of March is marked by a rise of the ratio of negative tweets to 0.55. This period was characterized by the French Prime Minister's announce that the lockdown was extended, high COVID-19 mortality and pressure on the health system. Another interesting point, which is reassuring for the validity of our classification algorithm can be found in Figure 2. Around the victory of France at the Football World Cup, a peak of positivity is clearly identifiable. Moreover, this shows that exogenous shocks can temporarily modify the modes of communication in the political sphere.
In the qualitative analysis of our predictions for the lockdown sample, it appears that positive tweets appeal to the notions of solidarity and support, and the thanking of elected officials and health professionals for handling the crisis (*cf.* Figure 3). In the negative tweets, MPs insist on the crisis and on lockdown, and they pay respect to victims and their families (*cf.* Figure 4). The qualitative evaluation of our performances confirms the satisfying results of the quantitative evaluation.
Furthermore, based on the lockdown subsample, we can identify sociodemographic and political factors that tend to influence the style of political discourse on Twitter. We refer here to the cross tables 1-4 and the multinomial logistic regression in table 5. It appears that gender has a role among MPs, with women being more informative / neutral than men in their tweets, particularly regarding negativity. Age seems to have an impact also, but in a less clear-cut way. A unit of increase of the number of mandates yields to more tweets being classified either as positive or negative compared to a neutral level of reference. From this, we speculate that an issue of legitimacy is at stake when producing discourse on Twitter: the more established you are (men as compared to women, older, more politically installed, *etc.*), the more you seem to convey strong sentiments. Finally, political factors play a great role in the style of communication: not belonging to the governmental majority increases substantially the tendency to produce negative tweets for French MPs, probably as criticisms of the governmental actions. This replicates at the level of the political groups, with a gradient of negativity associated to the ideological distance to the government.

Discussion. To improve our prediction algorithm, future work could elaborate on the pre-processing text-cleaning function. In our analyses, the # symbol is removed, but hashtags themselves are kept as such, which means that our word vectors contain non-spaced character strings. This would be relevant if these items were reused in the exact same form in all tweets. However, there might have variations, as the same meaning can be conveyed in different hashtags (*e.g.* "COVID19" or "Covid_19"). We tried to split non-spaced sentences by building a cost dictionary using Zipf's law and an array of French words, and writing a space-inferring function based on Python's `best_match()` function. However, this was inconclusive. Future research should try to adapt existing space-inferring functions to the French language.

[6]For more information on the OSCAR corpus, visit: https://oscar-corpus.com/
[7]For this reason, these further analyses are conducted in R which is more adequate than Python to produce statistics.
[8]All results are available on request.

References

Barrie, C. and Ho, J. (2021). academictwitteR: an R package to access the Twitter Academic Research Product Track v2 API endpoint. *Journal of Open Source Software*, 6(62):3272.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *arXiv:1607.04606 [cs]*. arXiv: 1607.04606.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Publisher: arXiv Version Number: 2.

Gillioz, A., Casas, J., Mugellini, E., and Khaled, O. A. (2020). Overview of the transformer-based models for NLP tasks. In Ganzha, M., Maciaszek, L. A., and Paprzycki, M., editors, *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020, Sofia, Bulgaria, September 6-9, 2020*, volume 21 of *Annals of Computer Science and Information Systems*, pages 179–183.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, V., Seddah, D., and Sagot, B. (2019). CamemBERT: a Tasty French Language Model. Publisher: arXiv Version Number: 3.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Appendix

All material - data and code - necessary to conduct the analyses are available here: https://github.com/mthsansu/MLNLP.

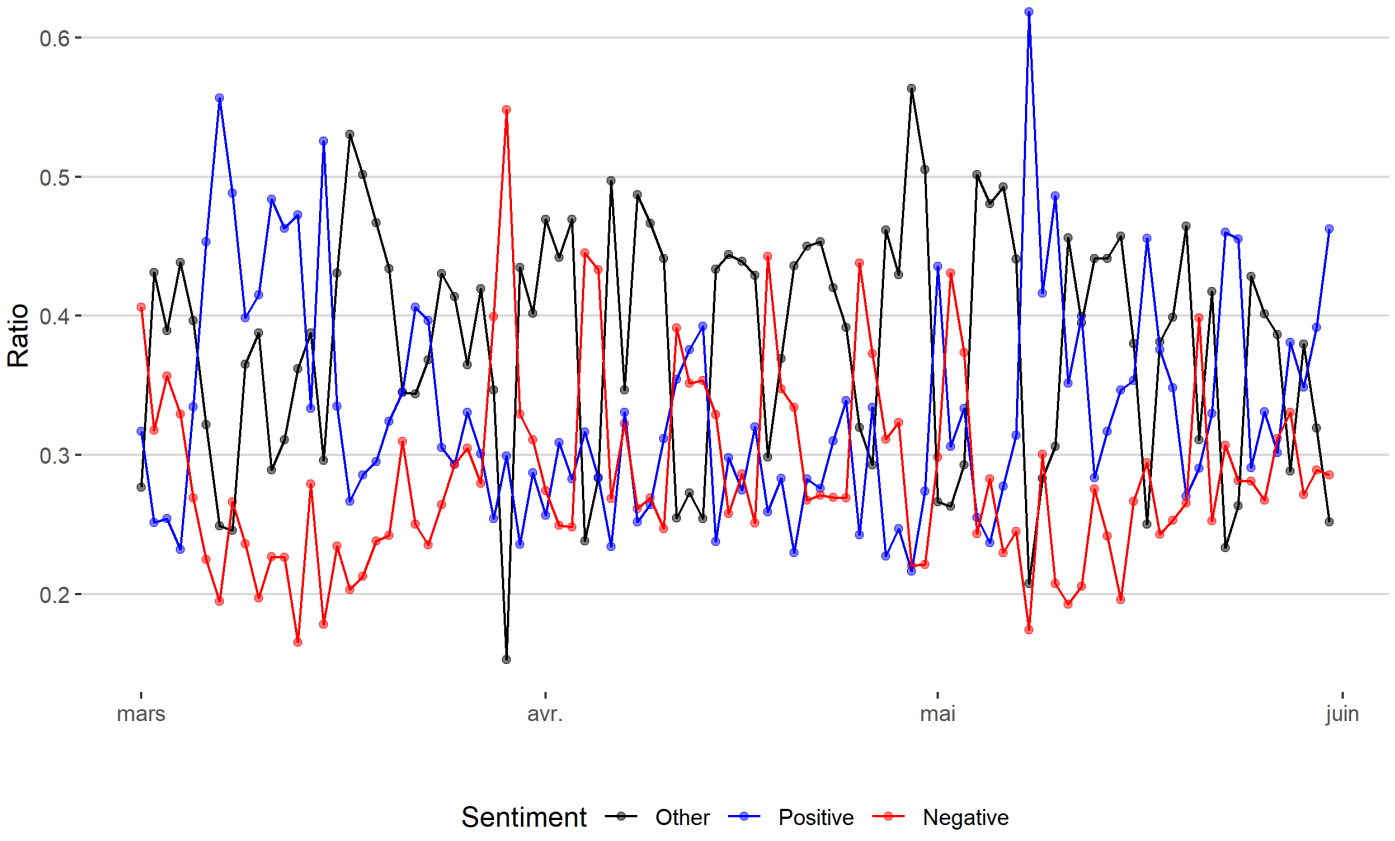

Figure 1: Evolution of ratios of tweets classified by sentiment around the first lockdown.

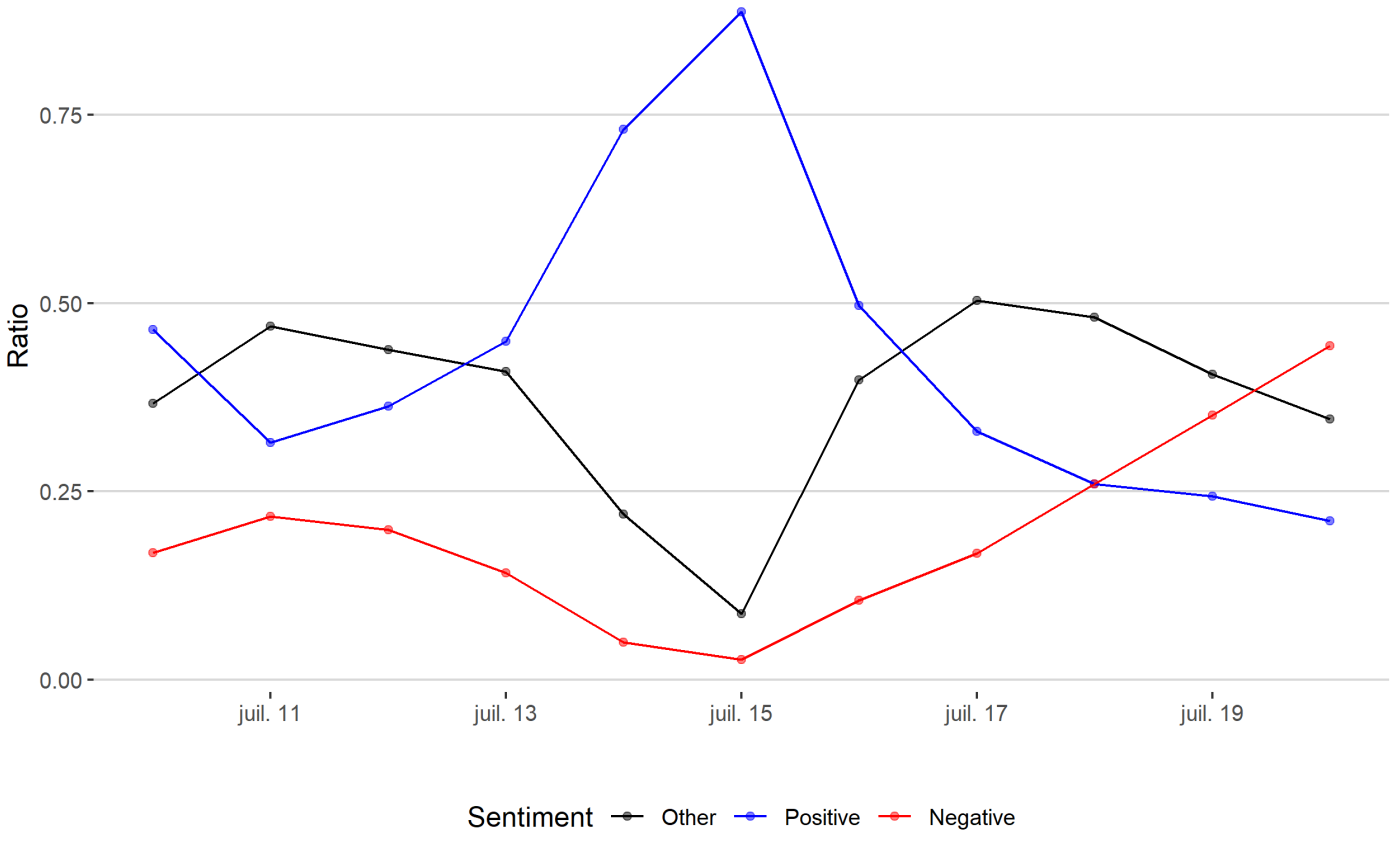

Figure 2: Evolution of ratios of tweets classified by sentiment around the Football World Cup.

Figure 3: Cloud of words for tweets classified as positive during the first lockdown.

Figure 4: Cloud of words for tweets classified as negative during the first lockdown.

	Other	Positive	Negative	Total
Women	0.43	0.35	0.22	12801
Men	0.37	0.31	0.32	19966
Total	0.39	0.33	0.28	32767

Table 1: Ratio of tweets published around the first lockdown by sentiment classified according to gender.

	Other	Positive	Negative	Total
GDR	0.31	0.21	0.48	673
LFI	0.25	0.16	0.59	3826
LR	0.29	0.32	0.38	4172
LREM	0.46	0.37	0.17	17740
MODEM	0.56	0.32	0.12	1964
NG	0.31	0.37	0.32	1101
NI	0.22	0.18	0.59	2036
UAI	0.38	0.42	0.20	1255
Total	0.39	0.33	0.28	32767

Table 2: Ratio of tweets published around the first lockdown by sentiment classified according to political group.

	Other	Positive	Negative	Total
24-34	0.33	0.37	0.30	4744
35-44	0.39	0.34	0.28	7735
45-54	0.39	0.32	0.28	11257
55-64	0.43	0.31	0.26	7468
65+	0.43	0.23	0.34	1563
Total	0.39	0.33	0.28	32767

Table 3: Ratio of tweets published around the first lockdown by sentiment classified according to age group.

	Other	Positive	Negative	Total
1	0.40	0.33	0.28	24594
2	0.39	0.33	0.28	7576
3	0.28	0.29	0.44	597
Total	0.39	0.33	0.28	32767

Table 4: Ratio of tweets published around the first lockdown by sentiment classified according to number of mandates.

Table 5: Multinomial Logistic Regression on tweet sentiment around the first lockdown.

	Dependent variable:	
	Positive	Negative
	(1)	(2)
Women	-0.050*	-0.267***
	(0.027)	(0.031)
Age	-0.018***	-0.014***
	(0.001)	(0.001)
Nb. mandates	0.063**	0.098***
	(0.030)	(0.033)
Group GDR	-0.270	0.234
	(0.203)	(0.200)
Group LFI	-0.433**	0.682***
	(0.180)	(0.184)
Group LR	0.223	0.168
	(0.172)	(0.180)
Group MODEM	-0.265***	-0.459***
	(0.054)	(0.075)
Group NG	0.365**	-0.042
	(0.186)	(0.194)
Group NI	-0.035	0.945***
	(0.168)	(0.175)
Group UAI	0.334***	0.324***
	(0.066)	(0.082)
Not in majority	0.098	1.073***
	(0.171)	(0.178)
Constant	0.564***	-0.341***
	(0.068)	(0.076)
Akaike Inf. Crit.	66,663.400	66,663.400
Note:	*p<0.1; **p<0.05; ***p<0.01	