Tweet like a future president
Machine Learning for Natural Language Processing 2022

Camille Francon
Ensae
camille.francon@ensae.fr

Khaled Larbi
Ensae
khaled.larbi@ensae.fr

Abstract

In this project, we wanted to compare two methods to produce tweets as five of the twelve candidates for the 2022 French presidential election. In order to generate these tweets, we mobilized two different algorithms: LSTM and GPT-2. The LSTM model is trained from scratch while the GPT-2 model is trained on a varied corpus of about 60 GB. The performances are evaluated quantitatively using the BLEU and qualitatively. Our results are available on Colab [1] or Github [2].

1 Problem Framing

The 2022 presidential campaign is a major event in French democratic life. For a few years now, it has not only taken place in rallies or on TV sets but also on social networks. More specifically, Twitter has become a place of confrontation with hashtags of the candidates' teams. How do candidates stand out on Twitter? Can we identify the styles of each candidate? Is it possible to simulate this style?

In order to answer these questions, we decided to try to reproduce tweets like the different presidential candidates. We restricted ourselves to five candidates that corresponded according to the polls of March 2022 (date when this project began) to the candidates gathering the most votes.

However, in order to obtain enough information, we considered for each candidate, four close accounts: often the official candidate account, the account for the candidacy and the accounts of two politically close. Some of the close individuals tweet occasionally: in some cases, we went far back in the timeline. The tweets were extracted using the Twitter API and the tweepy application. In total, nearly 38,000 tweets were extracted, trying to get as many for each candidate.

2 Experiments Protocol

To simulate the style of each of the five candidates, we will use two types of text generation models: LSTM and GPT-2. We have learnt a model for each candidat. However, some tweets contain only an emoji, an image or a link and therefore do not contain enough information to model the style of the candidates: we decided to set a minimum size of 60 characters for each tweet in order to be rich enough.

We also wanted to compare the addition of a pretrained model to a model learned from nothing: the LSTM model will be learned from nothing unlike the GPT-2 model learned on a large corpus in French (not specifically on political tweets).

3 Descriptive Statistics

The candidate for whom we have the most tweets is Marine Le Pen with 8481 tweets. Conversely, the candidate for whom we have the fewest tweets is Zemmour with 5605 tweets. This is significantly smaller than for the other 4 candidates.

For Macron, Le Pen, Mélenchon and Pécresse, The distribution of the number of characters is spread out to the right. This can be explained by the fact that the maximum number of characters on twitter is 280. The candidates therefore often use these 280 characters. For Zemmour, the distribution is more uniform with a mass around 280 all the same. The average number of characters is almost identical for Le Pen, Macron, Mélenchon and Pécresse (205 characters). As for Zemmour, it is much lower: about 163 characters.

[1] https://colab.research.google.com/drive/1gSDT18h14hxE9IFE_kEPZ6-Zxzg6JP6A?usp=sharing

[2] https://github.com/khaledlarbi/ENSAE_NLP_Tweets

Coupled with the fact that we have fewer tweets for Zemmour, this means that we will have less data to train the model associated with Zemmour.

For all candidates except Macron, and to a lesser extent for Zemmour, their # and @ are often used. In addition, for all the candidates, the words "France" and "français" are very present. For the far-right candidates, Le Pen and Zemmour, these words are used to a very large extent. For example, for Zemmour, the word "France" is used almost 4 times more than the 3rd most used word and the word "français" 2 times more. For Macron these words are also widely used but to a slightly lesser extent. We also notice that all the candidates often write Macron, except himself. The words "sécurité" and "protéger" are frequent in Macron's tweets, as are the words "ensemble" and "europe". The 10 most frequents words for each candidates are available in the appendix.

4 Results

4.1 GPT-2

GPT-2 for Generative Pre-trained Transformer 2 is a network created by Open AI in 2019 based on transformer methods which outperforms recurrent networks in sequence generations in the last years. GPT-2 is known as a general-ppurpose learner : theses models can be used for several tasks such as answers questions, summarization etc. Canonical GPT-2 is a model containing almost 1.5 billion parameters. In our project, we was working with a smaller version which contained 124 millions parameters. Initially, we wanted to work with GPT-3 but it was not release in a free access. The pretrained version of GPT-2 was from (Louis, 2020) who learnt it on a general French dataset (CommonCrawl, NewsCrawl, ...).

Good results when tweet begin by a mention to another user Here, we generate tweets with a initial seed which is "@JLMelenchon" as in figure 1.

We can notice that except for Macron, results seems realistics. Nevertheless, models don't seem to learn who is @JLMelenchon as we can see in the Pécresse tweets. Futhermore, we can notice that the pretrained model didn't provide a political tweet : then the fine-tunning is usefull here.

Bad results when a initial seed finishes by a non-alphanumerical character Here, we generate tweets with a initial seed which is "écologie " as in figure 2. We can note that sentences generated are not making sense because of extra space. When we removed it, we got better results.

4.2 LSTM

Long short-term memory (LSTM) is an architecture of recurrent neural network. This is often used in the context of text generation. Our model is composed of an LSTM layer and 3 fully connected layers.We use the architecture of this article (Musser, 2020). As input, we put 40 characters of a tweet and the expected output is the next character in the tweet. We obtain more than 1 million data for all the candidates except for Zemmour (it is almost 690 000). Thus, we trained the models on 5 epochs except that of Zemmour which was trained on 8 epochs.To generate tweet, we use a tool named temperature. It helps us not get the same text generated every time. if temperature is low, probabilities are more "extrem" and texts are similar to trained data. If temperature is high, probabilities are smoothed and texts are more creative. However, too high temperature can provoke nonsense.

Results : words that exist but sentences mean little Here, we generate tweets with a initial seed which is "concernant la guerre en ukraine,je pense". The temperature is set to 0.5.

We can see in figure 3 that the words and certain groups of words make sense but not the tweets as a whole. We recognize a few characteristics words or expressions of candidates like @vpecresse in these exemples.

5 Discussion/Conclusion

In conclusion, the results are better with GPT-2 than with LSTM. Indeed, the LSTM models generate existing words, but tweets often have no meaning. For GPT-2 models, tweets often make much more sense. Also, for LSTM models, the "base" must be at least 40 characters, which is not the case for GPT-2. We have not found how to quantitatively evaluate our models. We were interested in BLUE but the model outputs were too different from the reference texts. Training time for LSTM models is much longer (about 5h) than for GPT-2 (about 10min). We believe that a longer training time for LSTM could lead to better results.

MLP : @JLMelenchon à l' UMP, et au parti socialiste à la primaire de la gauche et du centre ", avait alors annoncé le numéro deux du parti.Le numéro trois avait annoncé en

Macron : @JLMelenchon. Ce qui m' a donné beaucoup de peine :) (Merci pour tous vos messages! Par ailleurs, le recours à une assurance responsabilité civile professionnelle pour le salarié qui vient

Melenchon : @JLMelenchon, je ne suis pas un crétin, je suis un humain! L' autre grand événement de la journée, c' est la fête du sport avec la présence de l
Zemmour : @JLMelenchon : " Moi-même, j' ai voté Macron " (vidéo) Marine Le Pen : " Le Pen, c' est la trahison des électeurs " (vidéo) Le

Pécresse : @JLMelenchon est le président du Parti, mais n' ai pas le soutien de ma famille politique. Il est donc très difficile de la définir en ces termes, parce que ce serait trop

Pretrained : @JLMelenchon qui n' en pouvait plus de se battre (oui, c' est vrai, c' est une idée que j' ai depuis très longtemps). Cette petite histoire en

Figure 1: Initial seed : @JMelenchon

MLP : écologie ui permet de retrouver le chemin de la connaissance de notre propre environnement, et d e trouver une solution adaptée aux évolutions climatiques. En ce temps-là, la famille ne voulait pas de moi

macron : écologie ossifiées : les chercheurs de l' Institut des Sciences de l' environnement et du CNRS ont découvert que l' acide biliaire peut entraîner des lésions cérébrales ou neurologiques ; en résumé,

melenchon : écologie ^ La ville de Lommel ^ Anvers ^ La ville de Caudry ^ Valenciennes ^ Roubaix ^ Lille ^ Va lenciennes

zemmour : écologie iteraliste) ont été particulièrement difficiles à appréhender parce que le mouvement de la Terre, malgré ses bouleversements, n' a jamais été aussi bien vu. C' est une maison de trois

pécresse : écologie □□ம□ஞ□□) est une série télévisée d' animation américaine en 2 épisodes créée par et diffusée

pretrained : écologie uro-génitale et les femmes enceintes avaient accès à leur traitement pour augmente r l' efficacité de leur traitement par l' atazanavir (Ataur) et l'

Figure 2: Initial seed : ecologie

Group of candidats	Accounts
Le Pen	@J_Bardella
	@JulienOdoul
	@MLP_officiel
	@RNational_off
Macron	@avecVous
	@CCastaner
	@EmmanuelMacron
	@StanGuerini
Melenchon	@AQuatennens
	@Francois_Ruffin
	@jlmelenchon
	@melenchon_2022
Pécresse	@avecValerie
	@BrunoRetailleau
	@othmannasrou
	@vpecresse
Zemmour	@DamienRieu
	@G_Peltier
	@reconquete_off
	@ZemmourEric

Table 1: Accounts used to extract tweets

Le Pen : concernant la guerre en ukraine,je pense de passer les français de la france et pour les pouvoirs publics pour rendre leur argent aux français leur pa

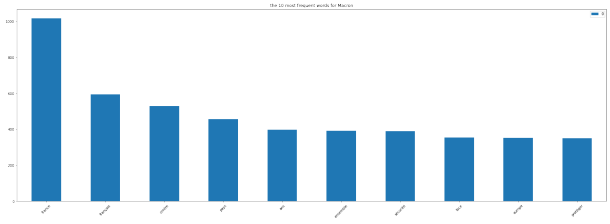
Mélenchon : concernant la guerre en ukraine,je pense que le chaos de chargé de la ministre de la transition sociale et son grand fortune de l'#unionpopulaire. #m

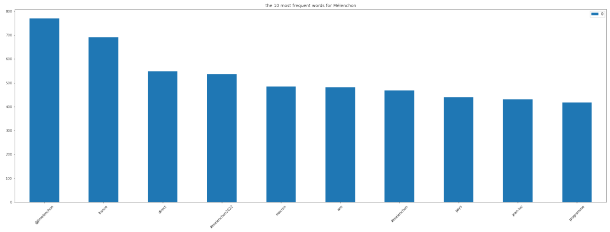
Pécresse : concernant la guerre en ukraine,je pense à la france en soutenant l'explosion de @vpecresse pour les français qui professent cette confrontation de la

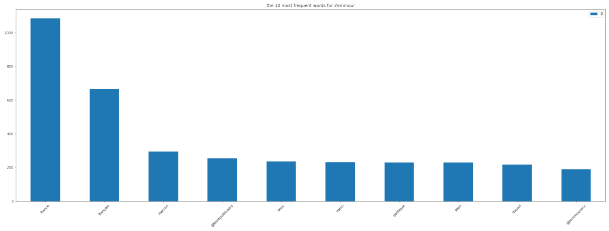
Macron : concernant la guerre en ukraine,je pense à sa vie pour la france et la sécurité et secours et son action de paris pour la france. avec son exemple de

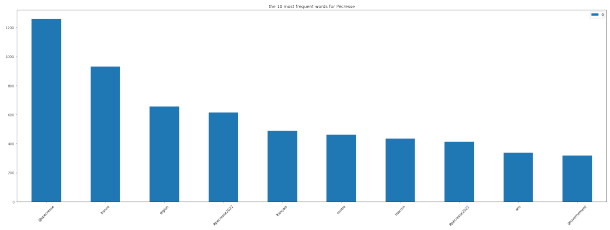
Zemmour : concernant la guerre en ukraine,je pense de proposer leur assistanat dans l'espace public, c'est le projet de l'esprit de l'immigration et de la franc

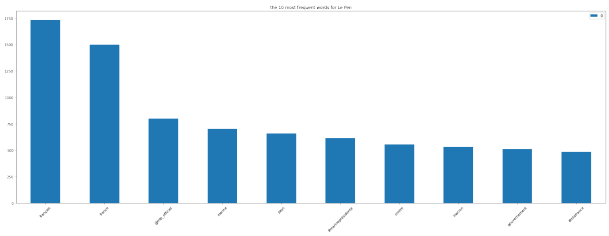
Figure 3: Initial seed : concernant la guerre en ukraine,je pense











References

Antoine Louis. 2020. BelGPT-2: a GPT-2 model pre-trained on French corpora. https://github.com/antoiloui/belgpt2.

Mikian Musser. 2020. Predicting Trump's Tweets With A Recurrent Neural Network.