Détection automatique du profil psychologique MBTI

Pour notre code et détails :
https://github.com/Yidi-Huang/Modelisation_MBTI

Kexin Gui
Yidi Huang

Vue d'ensemble du MBTI

- **Qu'est-ce que le MBTI ?**

  Le MBTI (Myers-Briggs Type Indicator) est un outil d'évaluation psychologique destiné à mesurer et à catégoriser les préférences comportementales et de personnalité des individus.

- **Origines et histoire**

  Basé sur la théorie des types psychologiques de Carl Jung, le MBTI a été développé par Isabel Myers et Katharine Briggs au milieu du 20e siècle.

- **Objectifs et applications**

  Utilisé dans le développement personnel, la gestion des ressources humaines, le conseil, l'orientation professionnelle, la psychothérapie, et l'éducation.

Les quatre dimensions du MBTI

https://www.16personalities.com/fr

- **Extraversion (E) vs. Introversion (I)**

  Extraversion: Orienté vers l'extérieur, trouve de l'énergie dans l'interaction avec les autres.

  Introversion: Orienté vers l'intérieur, trouve de l'énergie dans les réflexions personnelles.

- **Sensation (S) vs. Intuition (N)**

  Sensation: Privilégie les informations concrètes et actuelles, se fie à l'expérience.

  Intuition: Privilégie les possibilités, les concepts, se fie à l'inspiration.

- **Pensée (T) vs. Sentiment (F)**

  Pensée: Prend des décisions basées sur la logique objective et les principes universels.

  Sentiment: Prend des décisions basées sur les valeurs personnelles et l'harmonie sociale.

- **Jugement (J) vs. Perception (P)**

  Jugement: Préfère un mode de vie structuré avec des décisions prises.

  Perception: Préfère un mode de vie flexible avec des options ouvertes.

Problématique :

Est-ce que le modèle linguistique anglais de l'article est généralisable sur notre corpus français pour détecter MBTI à partir de l'analyse textuelle ?

-Est-ce que les données Twitter ne contiennent pas assez de signaux pour la détection de la personnalité MBTI ?

-Est-ce que les données textuelles ne correspondent pas bien aux scores de personnalité MBTI issus de questionnaires ?

-Selon Plank, B., & Hovy, D. (2015). *Personality traits on Twitter—or—how to get 1,500 personality tests in a week* pour un module, la suppression des mots vides nuit à la performance.

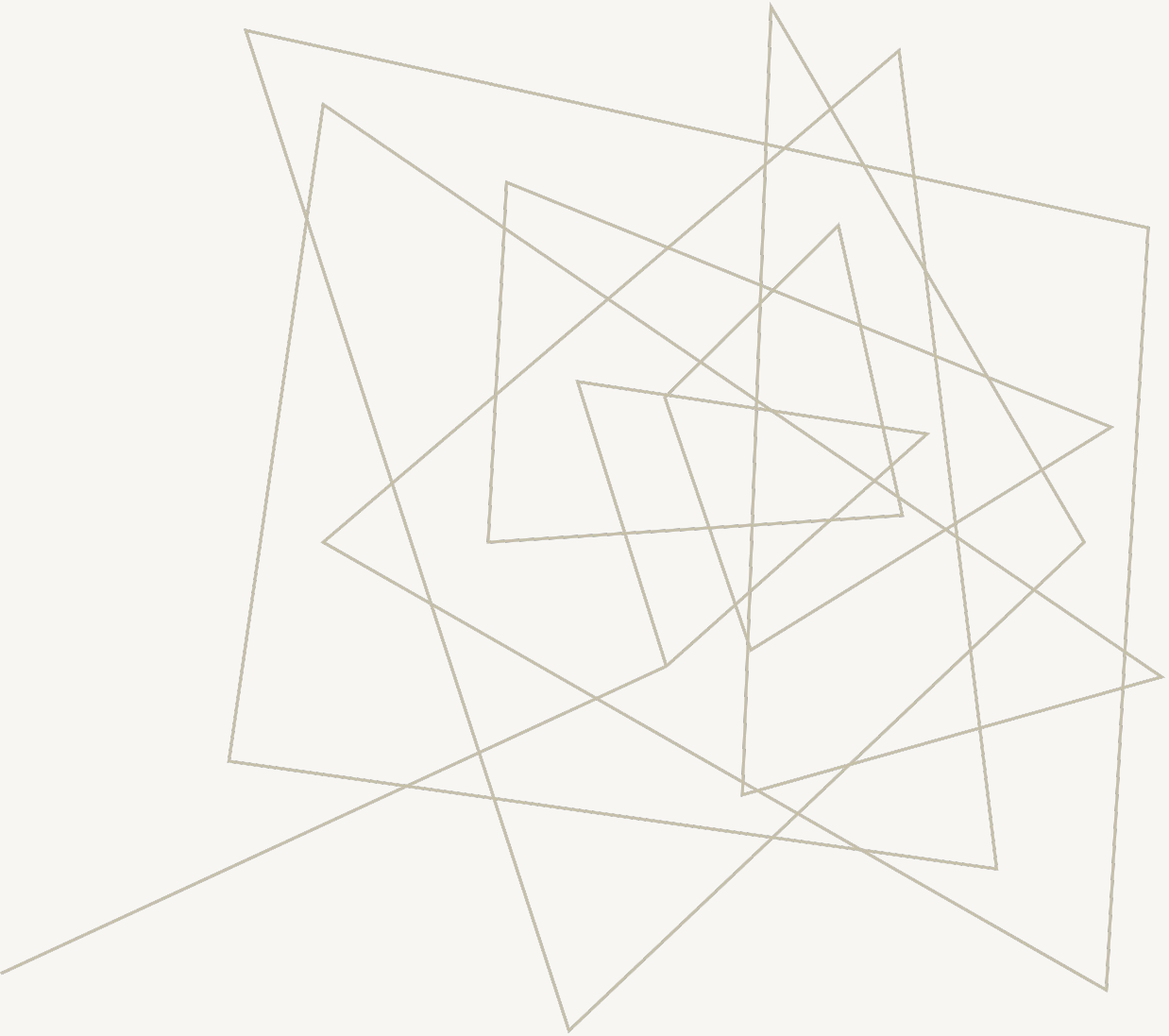
Plan

1 Synthèse de l’article

2 Construction du corpus

3 Reproduction du modèle linguistique

4 Utilisation du modèle



Synthèse de l’article

Why Is MBTI Personality Detection from Texts a Difficult Task? 2021
Sanja Štajner, Seren Yenikent

Contribution de l'article

Jeu de données

construire un nouveau jeu de données

Guide d'annotation

guide d'annotation pour l'analyse linguistique

Évaluation des hypothèses et des difficultés

Résultats et difficultés

Synthèse de l’article

- Jeu de données

MBTI-Twitter

- 6 échantillons aléatoirement sélectionnés pour chacun des 16 types MBTI → 96 au total
- conservé que les 10 premiers tweets (taille)

MBTI-MTurk

- Tâche d'intelligence humaine (HIT) Plateforme de crowdsourcing Amazon Mechanical Turk (MTurk)
- au moins 300 caractères
- 96 échantillons équilibrés

1. You might have done an MBTI personality test in the past. If you did, and you know the MBTI personality type you obtained, please write it here:
2. What is your favourite type of vacation and why?
3. Which are your favourite hobbies and why?

Synthèse de l'article

- Guide d'annotation

2 annotateurs : PhD en psychologie, PhD en linguistique informatique

Extravert	Introvert
Mention of new people (e.g. *crowd*, *strangers*)	Mention of closer people rather than any group of people (e.g. *husband*, *family*)
Mention of social activities and events that contain interaction with other groups of people (e.g. *party*, *dancing*, *couchsurfing*)	Mention of individual activities or activities that can be done without interaction with other people (e.g. *by myself*, *spending time at home*)
Mention of outside world and vibrant places (e.g. *bars*, *restaurants*)	Mention of inner world, and calm and quiet places (e.g. *home*, *museum*)
We references	*I* references
Use of intensifiers and exclamation marks	Hedging
More assertive, positive, enthusiastic arguments	Less assertive arguments

Table 1: Linguistic signals of extraversion and introversion.

Synthèse de l'article

- Guide d'annotation

Sensing	Intuitive
Technical, object-based and hands-on hobbies	Inspirational and imaginative hobbies (e.g. *creating*, *exploring*)
Facts and real cases (e.g. *documentary*, *diary*)	Abstraction rather than facts (e.g. *sci-fi*, *cartoons*)
Details and examples (more adjectives and adverbs to provide details, use of the words *example*, *for instance*)	Main ideas rather than details
Needs to use the 5 senses	Needs to focus on the bigger picture
Puzzles, model planes, crafts, carving, rowing, sailing, diving, rock climbing, etc.	Painting, music, dancing, poetry, chess, literature, arts, martial arts, yoga, meditation, etc.
Simplified and straightforward writing style (short sentences)	Complex writing style (long sentences)
Clear and concise writing style	Artistic, longer, more words

Table 2: Linguistic signals of sensing and intuition.

Synthèse de l'article

- Guide d'annotation

Thinking	Feeling
Logical reasoning for their actions and choices (e.g. *reading books for learning*)	Emotional reasoning for their actions and choices (e.g. *reading books for gateway feeling*)
Mention of opinions, ideas, comparisons	Mention of people, values, feelings
Direct (e.g. *reading is nice*)	Tactful, indirect (e.g. *reading feels nice*)

Table 3: Linguistic signals of thinking and feeling.

Synthèse de l'article

- Guide d'annotation

Judging	Perceiving
Holidays that include planning such as ski holidays, city tours etc. (e.g. *tour*, *pass*, *ticket*, *reservation*)	Spontaneous holidays such as going to the beach, a new city etc. (e.g. *flexible*, *spontaneous*)
Decisive, planful, organized (e.g. *plan*, *schedule*, *followed by*)	Curiosity, anticipation of change, and spontaneity
Organizers of the plans (e.g. *invite*, *organize*)	Followers of the plans (e.g. *join*, *tag along*)
Warranty (e.g. *insurance*, *make sure*)	Autonomy and impulsiveness (e.g. *suddenly*, *out of the blue*, *last minute*)
Past tense or present perfect tense	Present simple tense
Formal and structured writing style with grammatical rules followed as much as possible (e.g. *I like ski holidays and sometimes prefer city tours.*)	Informal writing style with grammar mistakes (e.g. *I like going to the beach. Also, do art sometimes.*)

Table 4: Linguistic signals of judging and perceiving.

Synthèse de l'article

- Évaluation de deux hypothèses et des difficultés

Statistic	MBTI-Twitter				MBTI-MTurk			
	EI	SN	TF	JP	EI	SN	TF	JP
Both annotators confident	53	**30**	43	**15**	62	**54**	38	**69**
Annotators agree	**46**	*81*	61	**50**	**100**	*69*	62	**78**
Annotator A agrees with the gold label	77	64	64	53	78	54	77	44
Annotator B agrees with the gold label	47	44	54	47	60	54	85	42
Both annotators agree with the gold label	77	54	57	50	75	62	54	43

Hypothèse 1 : Les messages Twitter ne contiennent pas toujours des signaux linguistiques permettant la détection de la personnalité.

Hypothèse 2 : Les données textuelles ne sont pas en adéquation avec les scores de personnalité MBTI obtenus à partir de questionnaires.

Construction du corpus

Construction du corpus

- Corpus de questionnaire et Corpus de Twitter

1.Vous avez peut-être déjà passé un **test de personnalité MBTI**. Si c'est le cas et que vous connaissez votre type de personnalité MBTI, veuillez le choisir ici:
(Si vous ne savez pas votre MBTI, vous pouvez faire le test ici https://mistypeinvestigator.com/test/v1)

2.Quel est votre **type de vacances préféré** (que voulez-vous faire pendant les vacances) et **pourquoi** ?
Veuillez rédiger une réponse d'environ 200 caractères.
Peut-être un peu long, mais s'il vous plaît cela va nous aider beaucoup ! :) Merci 💙

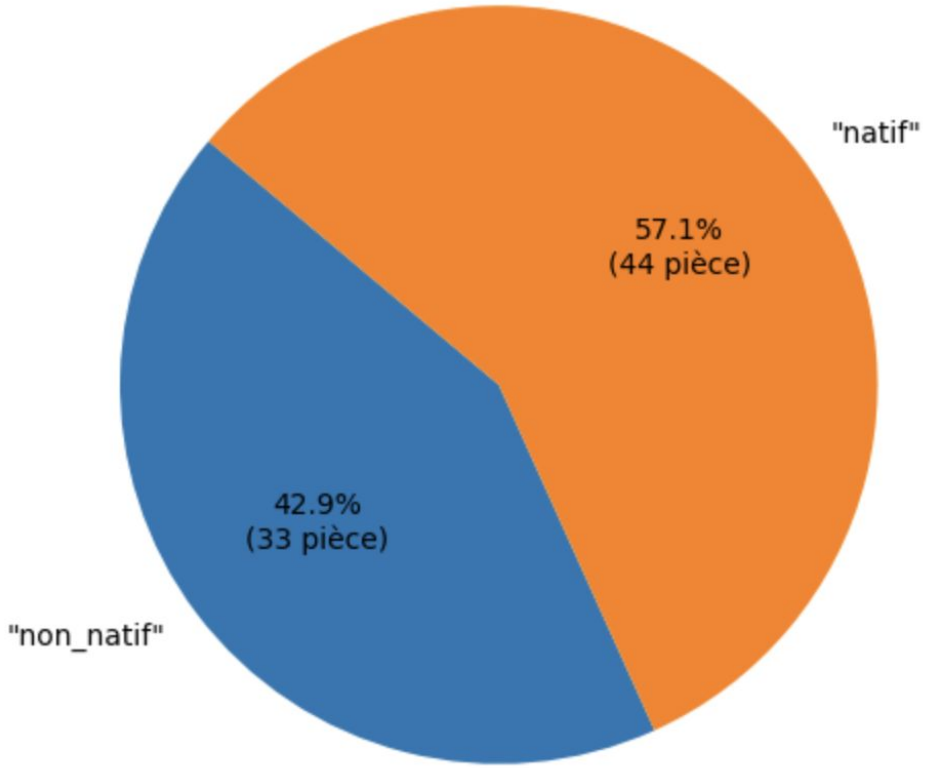
Long answer text

3.Quels sont vos **passe-temps préférés** et **pourquoi** ? *
Veuillez rédiger une réponse d'environ 200 caractères.
Peut-être un peu long, mais s'il vous plaît cela va nous aider beaucoup ! :) Merci 💗

Construction du corpus

- Corpus de questionnaire et Corpus de Twitter

résultat de questionnaire:
65 échantilons utiles.



résultat de Twitter :

16 types*4 → 64 échantillon

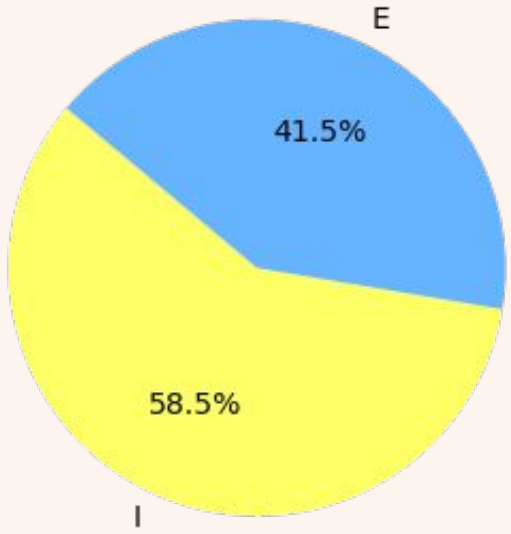
Reproduction du modèle

Reproduction du modèle linguistique

- Aperçu du corpus Twitter vs questionnaire

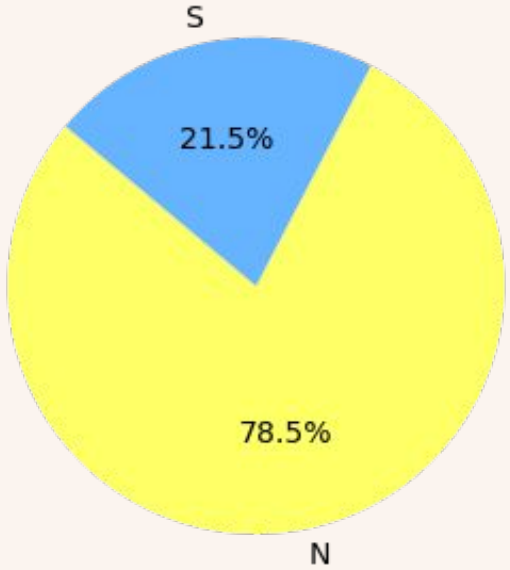
Total : 64 tweets regroupés vs 65 réponses

Repartition de I, E dans le corpus



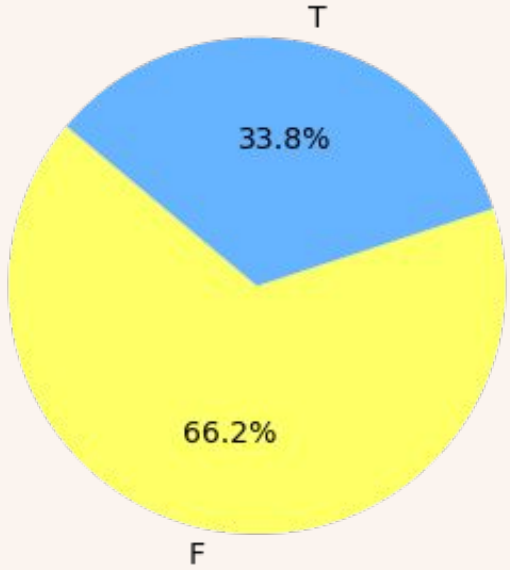
Catégorie	Fréquence
I	38
E	27

Repartition de N, S dans le corpus



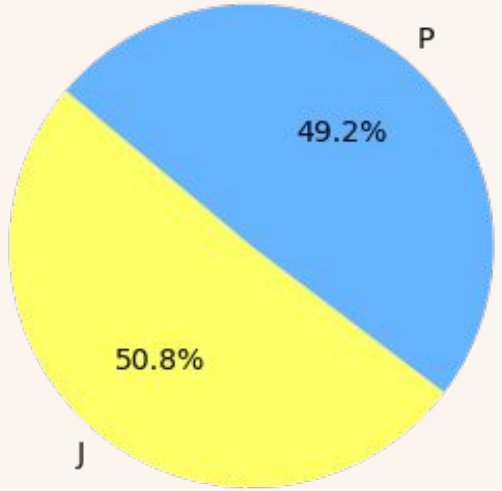
Catégorie	Fréquence
N	51
S	14

Repartition de F, T dans le corpus



Catégorie	Fréquence
F	43
T	22

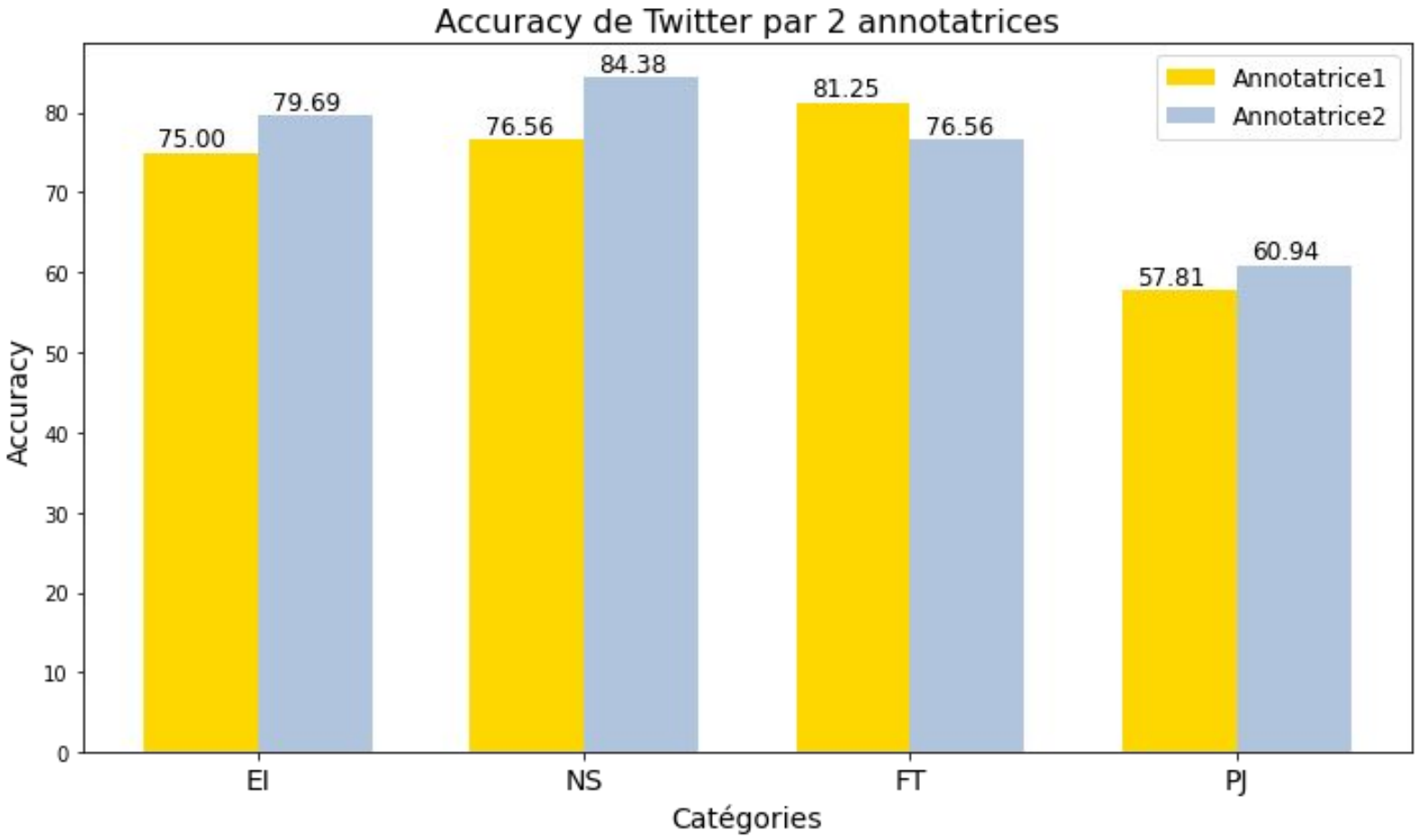
Repartition de J, P dans le corpus



Catégorie	Fréquence
J	33
P	32

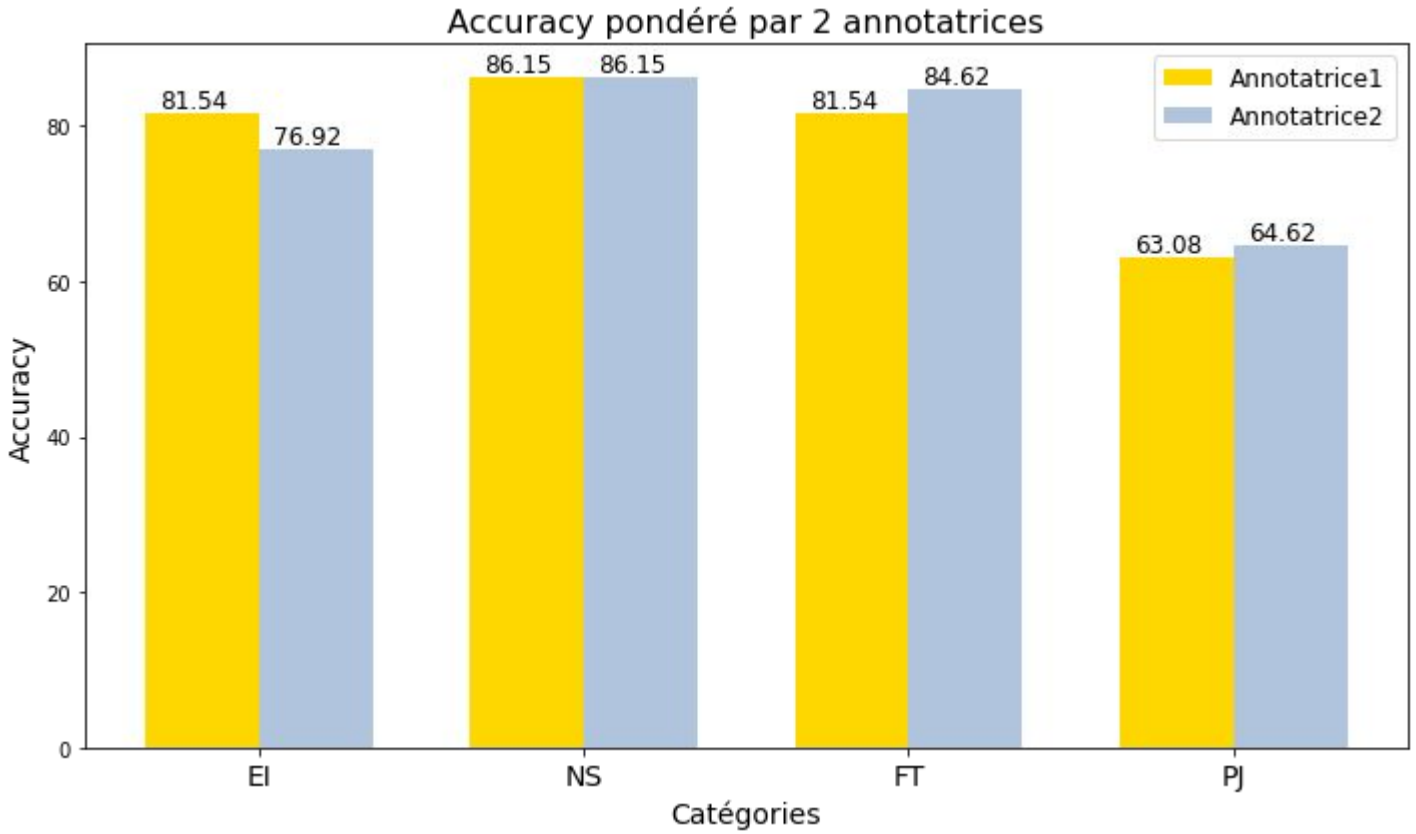
Reproduction du modèle linguistique

- Accuracy pour le corpus Twitter



Reproduction du modèle linguistique

- Accuracy pondéré pour le corpus questionnaire



$$Accuracy(EI)_{pondéré} = \frac{Accuracy(E)\times Corpus(E)+Accuracy(I)\times Corpus(I)}{Corpus(E)+Corpus(I)}$$

Reproduction du modèle linguistique

- Accord inter-annotateurs

Statistiques	Twitter				Questionnaire			
	EI	NS	FT	JP	EI	NS	FT	JP
2 annotateurs accord	39	42	41	29	45	50	46	30
A1 accord GOLD	48	49	52	37	53	56	53	41
A2 accord GOLD	51	54	49	39	50	56	55	42
2 annotateurs non-accord	4	3	3	17	7	3	3	13

Reproduction du modèle linguistique

- Accord inter-annotateurs pour le corpus Twitter

EI

		Annotatrice 1	Annotatrice 1
		Oui	Non
Annotatrice 2	Oui	39	9
Annotatrice 2	Non	12	4

Probabilité d'accord (Po) : 0.67
Probabilité d'accord simultané (Pe) : 0.65
kappa : 0.067

NS

		Annotatrice 1	Annotatrice 1
		Oui	Non
Annotatrice 2	Oui	42	7
Annotatrice 2	Non	12	3

Probabilité d'accord (Po) : 0.70
Probabilité d'accord simultané (Pe) : 0.68
kappa : 0.065

FT

		Annotatrice 1	Annotatrice 1
		Oui	Non
Annotatrice 2	Oui	41	11
Annotatrice 2	Non	9	3

Probabilité d'accord (Po) : 0.69
Probabilité d'accord simultané (Pe) : 0.68
kappa : 0.036

JP

		Annotatrice 1	Annotatrice 1
		Oui	Non
Annotatrice 2	Oui	29	8
Annotatrice 2	Non	10	17

Probabilité d'accord (Po) : 0.72
Probabilité d'accord simultané (Pe) : 0.52
kappa : 0.42

Reproduction du modèle linguistique

- Accord inter-annotateurs pour le corpus questionnaire

EI

		Annotatrice 1	Annotatrice 1
		Oui	Non
Annotatrice 2	Oui	45	8
Annotatrice 2	Non	5	7

Probabilité d'accord (Po) : 0.8
Probabilité d'accord simultané (Pe) : 0.67
kappa : 0.39

NS

		Annotatrice 1	Annotatrice 1
		Oui	Non
Annotatrice 2	Oui	50	6
Annotatrice 2	Non	6	3

Probabilité d'accord (Po) : 0.82
Probabilité d'accord simultané (Pe) : 0.76
kappa : 0.23

FT

		Annotatrice 1	Annotatrice 1
		Oui	Non
Annotatrice 2	Oui	46	7
Annotatrice 2	Non	9	3

Probabilité d'accord (Po) : 0.753
Probabilité d'accord simultané (Pe) : 0.72
kappa : 0.13

JP

		Annotatrice 1	Annotatrice 1
		Oui	Non
Annotatrice 2	Oui	30	11
Annotatrice 2	Non	11	13

Probabilité d'accord (Po) : 0.66
Probabilité d'accord simultané (Pe) : 0.53
kappa : 0.27

Reproduction du modèle linguistique

- Précision, Rappel et F-score d'annotation Twitter

Matrice				
	EI	SN	TF	JP
Précision	0.77	0.81	0.68	0.61
Rappel	0.77	0.8	0.66	0.59
F-score	0.77	0.8	0.66	0.58

Reproduction du modèle linguistique

- Précision, Rappel et F-score d'annotation questionnaire

Matrice				
	EI	SN	TF	JP
Précision	**0.84**	0.8	0.81	0.57
Rappel	0.76	0.8	0.81	0.53
F-score	0.76	0.79	0.81	0.51

Reproduction du modèle linguistique

- Exemple typique 1

Vacances :
Pendant les vacances, j'aime **peindre à la maison**, ou me retrouver **avec trois ou cinq amis** pour aller dans des vides greniers, ou parfois aller chez un ami pour dîner et discuter. Si les vacances sont assez longues, je peux prévoir un voyage lointain, ce qui est toujours agréable, que ce soit **seule ou avec des amis et des proches**.

Passe-temps :
Mes activités préférées sont **la lecture** et l'exercice, car je pense que la lecture permet à mon esprit de faire de l'exercice et que l'exercice permet à mon corps de rester en bonne condition. Chaque soir, avant d'aller me coucher, je **prends un livre** électronique et je lis pendant un certain temps, parfois des livres économiques, parfois des romans, et s'il s'agit d'un livre de science et de technologie, je peux m'endormir en un rien de temps, et je n'ai même pas besoin de mélatonine. L'exercice physique m'aide également à passer une bonne nuit de sommeil et à me réveiller le lendemain plein d'énergie.

E ou I ?

Reproduction du modèle linguistique

- Exemple typique 2

Vacances :
aller dans des endroits / lieux visuellement beaux
visiter de grandes villes
essayer de se fondre dans la vie des locaux
chercher des souvenirs qui me permettront de me souvenir de ce voyage / vacances / expériences toute ma vie
vivre des expériences en testant de nouvelles choses (cuisine, activité etc)
faire / m'offrir des choses que je ne fais pas en temps normal (aller au spa, m'offrir des choses onéreuses ...)

tout ça pour créer une distance avec le quotidien, mon environnement habituel, **pouvoir déconnecter**

Passe-temps :
me promener dans Paris, faire les magasins, trouver de nouveaux concept stores, aller au cinéma, lire un livre, faire des mot-mêlés, faire du petsitting

J ou P ?

Reproduction du modèle linguistique

- Reculs du modèle

H1 H2 : D’accord

Le modèle de l’auteur qui présente certaines limitations.

Les réponses de certaines personnes appartiennent véritablement à la zone médiane du spectre, montrant des caractéristiques des deux types de traits

Concernant la culture : Les signaux linguistiques proposés par l'auteur ont des caractéristiques culturelles anglaises assez marquées.

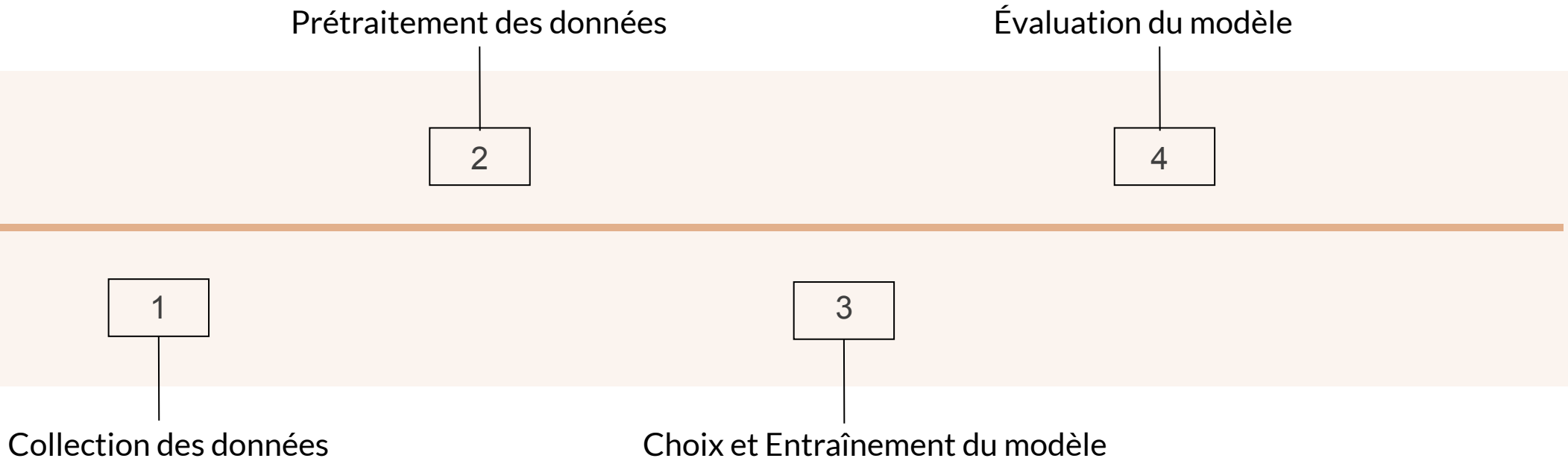
Est-ce qu'il existe différents « profils sociaux » au niveau culturel ?

- Dans les réponses “natif”, beaucoup de gens mentionnent des intérêts pour l'art, la musique, le cinéma, la littérature, la peinture, ce qui sont des loisirs assez communs présents dans les deux types de personnalités opposées.
- Le corpus des locuteurs natifs chinois reflète que les personnes d'origine culturelle chinoise ont tendance à être plus introverties.
- Les non-natifs, lorsqu'ils répondent aux questionnaires, se concentrent sur la grammaire et la conjugaison. Cela rend difficile la distinction dans la dimension JP.

Utilisation du modèle

Utilisation du modèle

- Démarches



Utilisation du modèle

- Prétraitement des données

1. Nettoyage

   - *MBTI mentionné, emoji, …*

2. Lemmatization

   - *Spacy*

3. Stop-words ?

notre
nous
nous-mêmes
nouveau
nouveaux
nul
…

```python
import os
import spacy

nlp = spacy.load("fr_core_news_md")

def lemmatize_text(text):
    doc = nlp(text)
    lemmatized_text = " ".join([token.lemma_ for token in doc])
    return lemmatized_text
```

```python
def nettoyage_and_rewrite_directory(directory_path):
    with open('stop-w_fr.txt', 'r', encoding='utf-8') as stop_words_file:
        stop_words = set(stop_words_file.read().splitlines())

    pattern = r'[\wÀ-ÖØ-öø-ÿ]+'

    for root, dirs, files in os.walk(directory_path):
        for file in files:
            file_path = os.path.join(root, file)

            if file_path.endswith(".txt"):
                with open(file_path, 'r', encoding='utf-8') as input_file:
                    text = input_file.read()

                cleaned_text = ""

                for word in re.findall(pattern, text.lower()):
                    if word not in stop_words:
                        cleaned_text += word + ' '

                with open(file_path, 'w', encoding='utf-8') as output_file:
                    output_file.write(cleaned_text)
```

Utilisation du modèle

- Choix et Entraînement du modèle

1. TfidfVectorizer
2. Word2Vec
3. CamemBERT ✓

```
camembert, tokenizer, weights = (ppb.CamembertModel, ppb.CamembertTokenizer, 'camembert-base')

tokenizer = tokenizer.from_pretrained(weights)
model = camembert.from_pretrained(weights)
```

✓ 1.3s

```
max_len = 0
for i,sent in enumerate(reviews):
    input_ids = tokenizer.encode(sent, add_special_tokens=True)
    if len(input_ids) > 512:
        print("annoying review at", i,"with length",
              len(input_ids))
    max_len = max(max_len, len(input_ids))

print('Max sentence length: ', max_len)
```

(code issu du : https://github.com/SidikiSidibe/classifieur-de-commentaires/tree/main)

Utilisation du modèle

- Evaluation camemBERT EI

Matrice				
	T-s	T-ns	Q-s	Q-ns
Accuracy	0.81	0.68	0.6	0.71
Précision	0.67	0.6	**0.86**	0.75
Rappel	0.8	0.86	0.43	0.43
F-score	0.73	0.71	0.57	0.55

Utilisation du modèle

- Evaluation camemBERT NS

Matrice				
	T-s	T-ns	Q-s	Q-ns
Accuracy	0.43	0.38	**0.81**	0.71
Précision	0.57	0.38	**0.81**	0
Rappel	0.4	1	**1**	0
F-score	0.47	0.55	**0.89**	0

Utilisation du modèle

- Evaluation camemBERT FT

Matrice				
	T-s	T-ns	Q-s	Q-ns
Accuracy	0.56	0.36	**0.7**	0.76
Précision	0.58	0	**0.7**	0
Rappel	0.78	0	1	0
F-score	0.67	0	0.82	0

Utilisation du modèle

- Evaluation camemBERT PJ

Matrice				
	T-s	T-ns	Q-s	Q-ns
Accuracy	0.5	0.5	**0.63**	0.41
Précision	0.46	0.5	**0.63**	0.45
Rappel	0.86	0.38	**1**	0.56
F-score	0.6	0.43	**0.77**	0.5

Conclusion

1.Impact de la conservation des mots vides sur les performances :

- La conservation des stopwords dans le Corpus de questionnaire a abouti à de meilleurs résultats de données.

2.Comparaison entre le Corpus de questionnaire et les textes Twitter :

- Globalement, les résultats obtenus avec le Corpus de questionnaire sont supérieurs à ceux des textes Twitter.

3.Différences de précision entre les différentes dimensions :

- Parmi les quatre dimensions du MBTI, l'exactitude est relativement plus élevée pour les dimensions Extraversion (E) et Introversion (I), tandis que la précision est la plus faible pour les dimensions Jugement (J) et Perception (P).

4.Le modèle Camembert est plus adapté à notre tâche.

Allez plus loin …

- Construire un corpus plus large.
- Établir des corpus dans différentes langues et mener des recherches dans différents contextes culturels, ce qui pourrait refléter les caractéristiques de personnalité des gens de différents pays et régions.

De l’approche de machine:

- Effectuer des ajustements supplémentaires des paramètres du modèle CamemBERT, ce qui pourrait optimiser les résultats.
- Essayer d'autres modèles de deep learning pré-entraînés.

Merci pour votre attention