

Choix de technologie

IAFA-tigable

Oleiwane Joe, Delmote Adrien, Meunier Mortimer, Bienassis Jules, Mouysset Martin

12/02/2024



M1 Informatique
S8
UE Management de projet

Résumé

Dans le cadre de notre sur la détection automatique de la subjectivité dans un texte de presse. Nous réalisons une revue des modèles existants afin d'effectuer le choix de l'architecture de celui que nous allons implémenter. Ce document synthétise toutes les informations pertinentes trouvées lors de la phase de recherche et argumente la direction actuelle de notre choix.

Table des matières

I	Introduction	3
II	LLMs “fine-tunable” (rapport des équipes de 2023)	4
III	BERT+CNN/LSTM	7
III.1	BERT+CNN	7
III.1.1	Architecture FakeBERT : BERT + CNN	7
III.1.2	Les résultats du modèle	8
III.2	BERT+LSTM	8
III.3	Limitations des modèles FakeBERT et BERT+LSTM	9
IV	LLMs généralisés	10
IV.1	Concernant l’utilisation des LLMs généralisés	10
IV.2	Collection de références	10
IV.3	Comparaison LLAMA	11
IV.3.1	Principales caractéristiques	11
IV.3.2	Principaux résultats	11
IV.3.3	Conclusion	12
IV.4	Comparaison Mistral et Mixtral	13
IV.4.1	Présentation des modèles	13
IV.4.2	Principaux résultats	14
IV.4.3	Conclusion	15
V	Conclusion : choix	16
VI	Références	17

I – Introduction

Le postulat de base indique la création de deux modèles de classification : un premier utilisant la technologie des LLM et un deuxième utilisant une architecture plus classique. Cependant, nos recherches ont révélé que les résultats des modèles LLM sur cette tâche sont supérieurs. Ainsi nous avons pu explorer différentes pistes :



- CheckThat 2023 : Utilisation de BERT et ses variantes (LLMs fine-tunables).
- BERT + LSTM et FakeBERT : BERT + CNN.
- Différent modèle de ‘LLMs généralisés’ (utilisés à travers des prompts).































Nous avons examiné les avantages et les inconvénients de chaque piste, en mettant l’accent sur l’importance de ces derniers quant à la détection de la subjectivité dans un texte.

II – LLMs “fine-tunable” (rapport des équipes de 2023)

Lors du “CheckThat! lab” de l’édition 2023 de l’initiative CLEF, une dizaine d’équipes ont proposé des modèles ayant pour but de classer des phrases comme étant soit objectives soit subjectives. Bien que ces équipes aient toutes utilisé des modèles différents, on peut noter de nombreuses similitudes. En effet, les équipes ont utilisé des variantes du modèle BERT (cf. “Table 3”)[7], qu’elles soient monolingues ou multilingues. Le modèle BERT permet en effet d’obtenir une représentation sémantique des mots et des phrases, très utile pour détecter la subjectivité. Ces modèles ont ensuite été “fine-tuned” sur les données spécifiques du projet de manière à apprendre à les classer. NB : ChatGPT n’est généralement pas utilisé lors de la classification. Une équipe a essayé de classer avec ChatGPT, mais a obtenu des résultats plus faibles que la baseline (cf. https://gitlab.com/checkthat_lab/clef2024-checkthat-lab/-/tree/main/task2/) pour l’ensemble des langues évaluées, nous n’utilisons donc pas leur architecture.

Table 3

Overview of the approaches to **Task 2**. The numbers in the language box refer to the position of the team in the official ranking; =part of the official submission; =considered in internal experiments.

Team		Languages						Models													Misc					
		Multilingual	Arabic	Dutch	German	English	Italian	Turkish	BERT	RoBERTa	XLNet	RoBERTa	GigaBERT	M-BERT	M-DeBERTa	S-BERT	SetFit	ChatGPT	GPT-3	BART	LSTM	Gradient Boosting	Multi-lingual training	Data augmentation	Feature Selection	Ensemble
Accenture	[35]		3	5	7	8	3	4																		
Awakened						10																				
DWReCo	[36]				4	1		2																		
ES-VRAI	[37]	-																								
Fraunhofer SIT	[38]				5	6																				
Gpachov	[39]					2																				
KUCST					4																					
NN	[40]	1	1	2	2	5	2	3																		
tarrekko		-	-	-	-	-	-	-																		
Thesis Titan	[41]	2	2	1	1	3	1	1																		
TOBB ETU	[42]	4	5	3	3	9	5	6																		
TUDublin	[43]					11	6																			

- Run submitted after the deadline.

cf. Overview of the CLEF-2023 CheckThat! Lab: Task 2 on Subjectivity in News Articles

Table 4

Results for the official submissions for the multilingual and for all six languages.

Team	F1	Team	F1	Team	F1
Multilingual		English		Italian	
1 NN [40]	81.97	- tarrekko *	78.19	1 Thesis Titan [41]	75.75
- tarrekko *	81.16	1 DWReCo [36]	78.18	- tarrekko *	71.61
2 Thesis Titan [41]	81.00	2 Gpachov [39]	77.34	2 NN [40]	71.01
- ES-VRAI [37]	77.96	3 Thesis Titan [41]	76.78	3 Accenture [35]	65.52
3 <i>baseline</i>	73.56	4 KUCST *	73.07	4 <i>baseline</i>	63.70
4 TOBB ETU [42]	66.62	5 NN [40]	72.84	5 TOBB ETU [42]	63.35
Arabic		6 Fraunhofer SIT [38]	72.72	6 TUDublin [43]	45.92
1 NN [40]	78.75	7 <i>baseline</i>	71.98	Turkish	
- tarrekko *	78.66	8 Accenture [35]	68.90	1 Thesis Titan [41]	89.94
2 Thesis Titan [41]	77.53	9 TOBB ETU [42]	63.46	- tarrekko *	87.01
3 Accenture [35]	72.53	10 Awakened *	60.41	2 DWReCo [36]	84.11
4 <i>baseline</i>	65.75	11 TUDublin [43]	40.32	3 NN [40]	81.21
5 TOBB ETU [42]	64.51	German		4 Accenture [35]	78.11
Dutch		1 Thesis Titan [41]	81.52	5 <i>baseline</i>	77.40
† 1 Thesis Titan [41]	81.43	2 NN [40]	74.13	6 TOBB ETU [42]	70.16
- tarrekko *	77.74	- tarrekko *	73.08		
2 NN [40]	75.57	3 TOBB ETU [42]	71.19		
3 TOBB ETU [42]	73.01	4 DWReCo [36]	69.82		
4 <i>baseline</i>	66.68	5 Fraunhofer_SIT [38]	68.39		
5 Accenture [35]	62.32	6 <i>baseline</i>	63.65		
		7 Accenture [35]	25.58		

- Run submitted after the deadline.

† Team involved in the preparation of the data.

*No working note submitted.

cf. Overview of the CLEF-2023 CheckThat! Lab: Task 2 on Subjectivity in News Articles

Du point de vue du volume des données, nous avons noté que la majorité des équipes ont considéré le label-bias : déséquilibre dans la quantité de données étiquetées subjectives et objectives comme un point de qualité essentiel qui affecte les résultats. Plusieurs techniques de "data augmentation" ont été utilisées comme par exemple :

- chatGPT pour générer une phrase de la classe opposée d'une phrase donnée (cf. Fraunhofer SIT at CheckThat! 2023: Can LLMs Be Used for Data Augmentation & Few-Shot Classification? Detecting Subjectivity in Text Using ChatGPT et TUDublin at CheckThat! 2023: ChatGPT for Data Augmentation).
- "under-sampling and over-sampling" : sous-échantillonnage sur la classe minoritaire et sur-échantillonnage sur la classe majoritaire des données. (cf. DWReCO at CheckThat! 2023: Enhancing Subjectivity Detection through Style-based Data Sampling).
- back-translation : un processus de rétro-traduction vers une langue temporaire (cf. Accenture at CheckThat! 2023: Impacts of Back-translation on Subjectivity Detection).
- data aggregation : processus qui réunit tous les dataSets en un seul multilingue pour augmenter la taille du set d'apprentissage. (cf. ES-VRAI at CheckThat! 2023: Enhancing Model Performance for Subjectivity Detection through Multilingual Data Aggregation).

De plus, Une équipe a analysé l'impact du style d'écriture générée sur la classification (cf. DWReCO at CheckThat! 2023: Enhancing Subjectivity Detection through Style-based Data Sampling) ils démontrent à partir de cela que l'augmentation de données en tenant compte du style d'écriture est plus avantageuse que de la paraphrase, sauf que le meilleur style choisi diffère d'une langue à une autre.

Si le choix d'utiliser un BERT fine-tuné est fait, il conviendra donc d'essayer les méthodes de traitement des données qui s'avèrent efficaces afin de mesurer les diverses performances en essayant plusieurs combinaisons de paramètres.

Ces architectures seront plus simples à mettre en œuvre grâce au nombre d'exemples et d'hyperparamètres cités dans les rapports de l'an dernier. Les équipes ont rencontré une légère difficulté quant à l'amélioration de leurs modèles à cause du manque de données. De plus, les meilleurs scores obtenus sont autour de 80.00 (en score F1, score prenant en compte les faux positifs et faux négatifs)(cf "Table 4")[7], il convient donc de s'assurer qu'il n'existe pas déjà une autre possibilité obtenant de meilleurs résultats.

III – BERT+CNN/LSTM

III.1 BERT+CNN

Dans le but de détecter les fake news, Rohit Kumar Kaliyar , Anurag Goswami et Pratik Narang nous présentent un modèle de LLM pouvant classifier si une news est véridique.[2].

Le modèle présenté combine à la fois une base d'encodage BERT relié en sortie à un réseau de neurones convolutionnel qui finit par classifier l'entrée donnée au modèle entraîné.

Le jeu de données utilisé se compose de deux fichiers train.csv et test.csv : un ensemble de données de test sans labels. Il s'agit d'une collection de fausses et vraies nouvelles propagées à l'époque de l'élection présidentielle générale américaine-2016.

Les chercheurs ont aussi testé d'autres modèles inclus dans leur rapport tel que ; En Deep Learning : CNN avec BERT, LSTM avec GloVe, LSTM avec BERT. En MachineLearning : Multinomial Naïve Bayes with GloVe (0.89 accuracy), K-Nearest Neighbors with GloVe, Decision Tree with GloVe word embedding (0.73 accuracy), Random Forest with GloVe.

III.1.1 Architecture FakeBERT : BERT + CNN

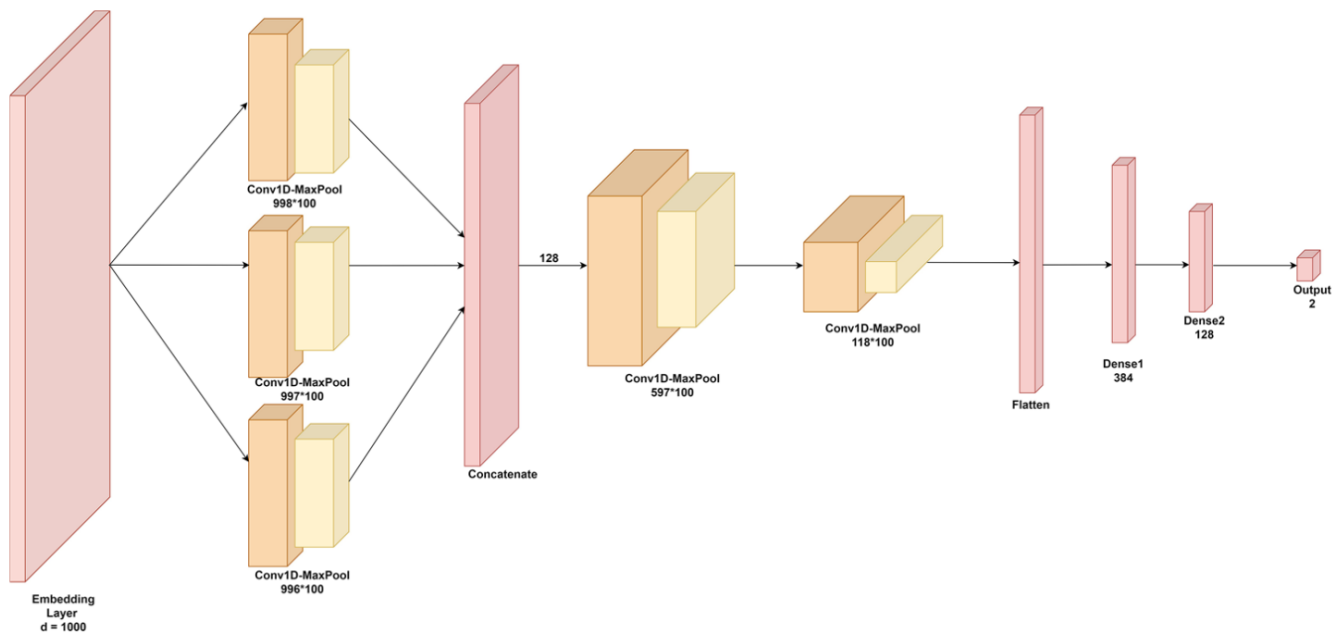


Fig. 5 FakeBERT model

(cf. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach)

Le FakeBERT se compose de 5 couches :

- Couche de convolution chargée d'améliorer la représentation sémantique des mots de longueur différente.
- Couche de mise en commun maximale chargée de réduire la taille de la représentation des mots afin d'optimiser le nombre de calculs.
- Couche d'aplanissement chargée de modifier le format des données.
- Deux couches denses chargées d'effectuer la classification finale.

III.1.2 Les résultats du modèle

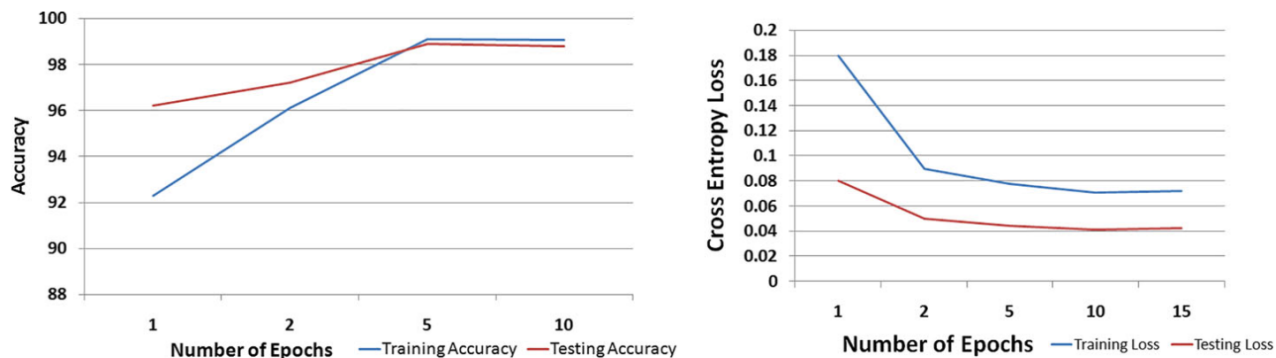


Fig. 8 Accuracy and cross entropy loss using FakeBERT

Précision atteinte : 0.989

Le graphe de précision montre que l'exactitude du modèle sur les données d'entraînement et de test augmente avec le nombre d'époques. Il semble que le modèle n'ait pas de surajustement (overfitting) significatif puisque la précision de test suit de près la précision d'entraînement. De même, le graphe de la loss indique une diminution typique au fur et à mesure que le nombre d'époques augmente, ce qui est attendu lors de l'entraînement d'un modèle neuronal. Encore une fois, le fait que la perte de test suive la perte d'entraînement indique que le modèle généralise bien et n'est pas surajusté.

Table 19 Confusion matrix for FakeBERT with BERT

	Predicted negative	Predicted positive
Actual negative	1045 (TN)	6 (FP)
Actual positive	17 (FN)	1012 (TP)

(cf. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach)

Ces résultats montrent un taux très élevé de vrais positifs et de vrais négatifs, ce qui indique une performance élevée dans la classification correcte des nouvelles. Les nombres relativement faibles de faux positifs et de faux négatifs montrent que le modèle est fiable et ne commet pas beaucoup d'erreurs de classification.

La précision de 0,989 indique que le modèle est très performant pour une tâche de classification, mais il faudrait réexaminer et potentiellement reconfigurer le modèle pour qu'il puisse distinguer correctement entre subjectivité et objectivité. Les caractéristiques linguistiques qui déterminent la subjectivité peuvent être subtiles et différer de celles utilisées pour juger de la véracité. Par exemple, le modèle devra être affiné pour reconnaître les opinions, l'ironie et le sarcasme, et les expressions discriminantes d'après la définition de la subjectivité de CheckThat On the Definition of Prescriptive Annotation Guidelines for Language-Agnostic Subjectivity Detection, plutôt que la présence de faits vérifiables.

III.2 BERT+LSTM

Un autre bon modèle, quoiqu'ayant montré des résultats légèrement plus faibles, est l'utilisation d'une couche de LSTM après les couches de convolution.

Les modèles de type BERT permettent une représentation précise de la sémantique de la phrase, tandis qu'une architecture de type LSTM (ou les deux) permet de classifier en utilisant des "patterns" assez complexes, nous semblant pertinent pour détecter la subjectivité, tel que le modèle BERT + LSTM suivant détectant le sarcasme, dont l'architecture est décrite comme suit :

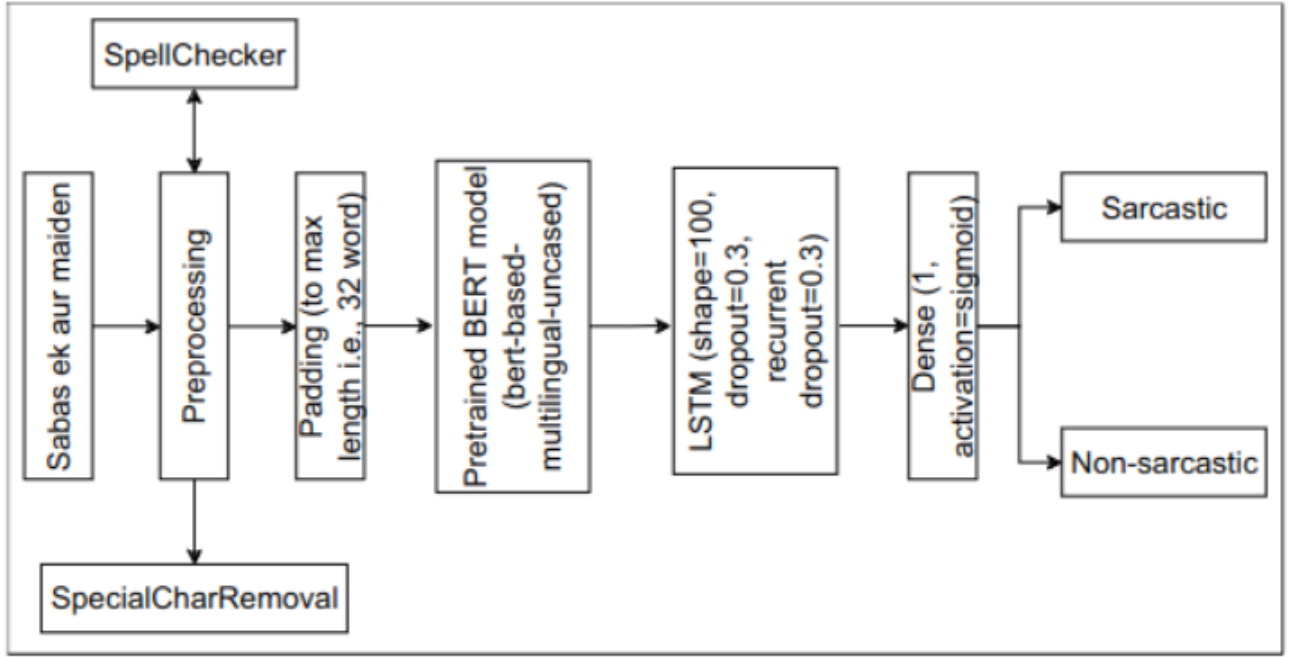


Fig. 3 The proposed BERT-LSTM network for sarcasm detection

(cf. BERT-LSTM model for sarcasm detection in code-mixed social media post)

Une couche de neurone dense s'occupe ensuite de faire la classification en deux classes. Nous supposons ici que la détection de sarcasme est proche de la détection de subjectivité. En effet, une phrase contenant du sarcasme est automatiquement considérée comme subjective. Sur le dataset entraîné, le modèle a obtenu un f1 score de 0.98, ce qui est excellent sur la tâche précise sur laquelle il a été entraîné.

III.3 Limitations des modèles FakeBERT et BERT+LSTM

Ces deux architectures, bien qu'efficaces, requièrent un grand nombre de données (le minimum observé pour un modèle monolingue est 6000 données distinctes). Cela peut être un problème dans le cadre de notre projet où le nombre de données est assez réduit (6000 toutes langues confondues, entre 800 et 1600 par langue). Ces modèles ayant été testé sur des ensembles de données monolingues, c'est un problème à résoudre si on veut utiliser ce type d'architecture.

IV.1 Concernant l'utilisation des LLMs généralisés

Contrairement aux autres pistes qui possèdent de la littérature concernant cette tâche particulière de classification de texte, nous n'avons pas trouvé d'information concernant l'utilisation de LLM généralisés comme Llama ou Mistral. Cela pourrait être un nouvel angle d'attaque du problème qui pourrait être pertinent mais demandera de développer en grande partie nous même la solution. On pourrait, par exemple, se baser sur un mistral fine-tuned déjà existant.

Dans le but d'explorer la piste des LLM généralisés dans la détection de textes subjectifs, nous allons devoir effectuer une comparaison afin de choisir celui qui semble le plus efficace.

IV.2 Collection de références

Une collection de références est un ensemble de tests standardisés utilisé pour évaluer et comparer les performances des modèles d'intelligence artificielle, en particulier dans le domaine de la compréhension du langage naturel.

Chaque benchmark est conçu pour tester des aspects spécifiques de la compréhension et du raisonnement d'un modèle d'IA. Par exemple :

Explication des benchmarks pour compréhension des résultats :

- BoolQ (Boolean Questions) : Un benchmark de compréhension de lecture où les modèles répondent à des questions fermées par oui ou non en se basant sur un passage de texte court.
- PIQA (Physical Interaction QA) : Teste la compréhension physique d'un modèle par des questions pratiques, demandant de prédire l'issue d'interactions physiques dans des scénarios donnés.
- SIQA (Social Interaction QA) : Un benchmark de raisonnement social qui pose des questions basées sur des scénarios de la vie quotidienne pour tester la compréhension des normes sociales et des conséquences des actions.
- HellaSwag : Un benchmark pour l'évaluation de la compréhension du langage modèle et la prédiction de fin d'histoires. Les modèles doivent choisir la continuation la plus plausible d'un scénario parmi plusieurs options.
- WinoGrande : Un test de raisonnement de bon sens à grande échelle qui améliore le WSC (Winograd Schema Challenge) original. Il évalue la capacité d'un modèle à résoudre des problèmes de désambiguïsation pronominale.
- ARC (AI2 Reasoning Challenge) Easy (ARC-e) et Challenge (ARC-c) : Un ensemble de questions de quiz scientifique de niveau scolaire. ARC-e contient des questions plus faciles, tandis que ARC-c comprend des questions jugées plus difficiles pour les modèles d'IA.
- OBQA (Open Book Question Answering) : Un benchmark de compréhension de lecture qui teste la capacité du modèle à répondre à des questions de connaissance générale sans accéder à une source externe d'information (comme si l'examen était "à livre ouvert").

De plus, nous pourrions utiliser la plateforme de Hugging Face comportant un classement des LLM open source par rapport à ces benchmarks. Ce classement est mis à jour chaque semaine et nous permet de récupérer des modèles modifiés.

IV.3 Comparaison LLaMA

Meta AI propose une gamme de modèles LLaMA (Large Language Model Meta AI) de traitement du langage naturel avec des architectures allant de 7B à 65B paramètres [1], utilisant un algorithme de tokenisation byte-pair encoding (BPE) basé sur le SentencePiece de Kudo et Richardson (2018). Les modèles ont été entraînés sur un ensemble de données opensources, offrant ainsi une variété de capacités adaptées à divers besoins en NLP.

IV.3.1 Principales caractéristiques

L'architecture de LLaMA a évolué par rapport à l'architecture originale Transformer voici ses modifications :

- LLaMA utilise une technique de pré-normalisation où la normalisation est appliquée à l'entrée de chaque sous-couche du transformateur au lieu de sa sortie. Ceci est fait pour améliorer la stabilité durant l'entraînement. La fonction de normalisation utilisée est RMSNorm, qui a été introduite par Zhang et Sennrich en 2019.
- Au lieu de l'activation ReLU habituellement utilisée dans les réseaux de neurones, LLaMA utilise la fonction d'activation SwiGLU, proposée par Shazeer en 2020. Cette fonction est connue pour améliorer les performances du modèle. LLaMA utilise une dimension de $2/3 \times 4d$ pour cette fonction d'activation, contrairement à PaLM qui utilise une dimension de $4d$.
- LLaMA élimine les embeddings positionnels absolus et utilise à la place des embeddings positionnels rotatifs (RoPE), qui ont été introduits par Su et al. en 2021. Ces embeddings sont ajoutés à chaque couche du réseau et permettent au modèle de mieux capturer la relation entre les positions dans les séquences de données.

IV.3.2 Principaux résultats

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	88.0	82.3	-	83.4	81.1	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	80.0	57.8	58.6
	65B	85.3	82.8	52.3	84.2	77.0	78.9	56.0	60.2

Table 3: Zero-shot performance on Common Sense Reasoning tasks.

(cf. LLaMA: Open and Efficient Foundation Language Models)

Tâches en Zero-shot et Few-shot :

- Zero-shot : Description textuelle de la tâche et exemple de test donnés, le modèle génère une réponse ou classe des réponses proposées.
- Few-shot : Quelques exemples de la tâche (entre 1 et 64) et un exemple de test sont donnés, le modèle génère la réponse ou classe différentes options.

(cf. LLaMA: Open and Efficient Foundation Language Models)

Performances sur les benchmarks :

LLaMA montre des performances compétitives sur une variété de benchmarks de raisonnement de bon sens, y compris BoolQ, PIQA, SIQA, HellaSwag, WinoGrande, ARC (facile et challenge) et

OpenBookQA. De plus, Sur des benchmarks de question-réponse en livre fermé comme Natural Questions et TriviaQA, LLaMA-65B atteint des performances de pointe en zero-shot et few-shot. Et pour finir, LLaMA-13B est compétitif avec GPT-3 et Chinchilla malgré sa taille plus petite, fonctionnant sur un seul GPU V100 pendant l'inférence.

Compréhension massive en multitâche :

		Humanities	STEM	Social Sciences	Other	Average
GPT-NeoX	20B	29.8	34.9	33.7	37.7	33.6
GPT-3	175B	40.8	36.7	50.4	48.8	43.9
Gopher	280B	56.2	47.4	71.9	66.1	60.0
Chinchilla	70B	63.6	54.9	79.3	73.9	67.5
PaLM	8B	25.6	23.8	24.1	27.8	25.4
	62B	59.5	41.9	62.7	55.8	53.7
	540B	77.0	55.6	81.0	69.6	69.3
LLaMA	7B	34.0	30.5	38.3	38.1	35.1
	13B	45.0	35.8	53.8	53.3	46.9
	33B	55.8	46.0	66.7	63.4	57.8
	65B	61.8	51.7	72.9	67.4	63.4

Table 9: Massive Multitask Language Understanding (MMLU). Five-shot accuracy.

(cf. LLaMA: Open and Efficient Foundation Language Models)

Sur le benchmark MMLU, LLaMA-65B est légèrement derrière Chinchilla et PaLM, ce qui pourrait être dû à l'utilisation limitée de livres et de documents académiques dans les données d'entraînement de LLaMA.

Point sur la toxicité, biais et vérité de réponses du modèle

Les modèles de langage, incluant LLaMA, peuvent produire du langage toxique et présenter des biais, notamment en termes de genre, âge, et religion. Des évaluations telles que RealToxicityPrompts, CrowS-Pairs, et WinoGender révèlent une augmentation de la toxicité avec la taille du modèle et des biais spécifiques, bien que LLaMA surpasse légèrement d'autres modèles. Malgré des scores plus élevés dans TruthfulQA par rapport à GPT-3, indiquant une meilleure adéquation à la vérité, les réponses de LLaMA restent mitigées.

IV.3.3 Conclusion

Pour conclure, LLaMA est un modèle de langage puissant et polyvalent avec des améliorations architecturales significatives et des optimisations efficaces. Cependant, il fait face à des défis liés à la toxicité, aux biais, et à la qualité des réponses. De plus, il est moins efficace que certains modèles dans la compréhension massive en multitâche. Nous verrons à travers l'étude suivante sur le LLM Mixtral comment répondre à ces manques, pour répondre aux besoins spécifiques d'applications comme la détection de la subjectivité dans les FakeNews.

IV.4 Comparaison Mistral et Mixtral

La société MistralAI propose deux modèles Open sources, Mistral 7B et Mixtral 8x7B dans leur papier de présentation, ces modèles sont comparés aux modèles Llama et GPT.

IV.4.1 Présentation des modèles

Mistral 7B est un modèle basé sur l'architecture Transformers comme les modèles Llama, mais avec quelques modifications.

- Un système d'attention par fenêtre glissante, au lieu que tous les tokens précédent soient conservés, l'attention n'est gardée que sur un nombre fixe de tokens qui s'enchaînent au fur et à mesure que le traitement se réalise. Cela augmente la vitesse de traitement et donc réduit la latence pour un traitement du même nombre de tokens tout en conservant l'influence de tous les tokens pour réaliser les prédictions.
- Le nombre fixe de tokens permet d'implémenter une mémoire tampon roulante. Au lieu de devoir conserver tous les tokens précédent entre 2 séparateurs, et donc d'avoir un cache de taille croissant tout le long de la prédiction, ici le cache atteint une taille maximale et est écrasé constamment au long de la prédiction. Cela permet de réduire la quantité de mémoire utilisée sans toucher à la qualité des prédictions.
- La prédiction des tokens se fait en séquence, mais le prompt est connu dans sa totalité dès le début, il est donc découpé en morceau ce qui permet de définir la taille de la fenêtre à la taille des morceaux et de pré-remplir la mémoire tampon par morceaux plutôt que mot à mot.

Mixtral 8x7B est un modèle d'un nouveau type, Sparse Mixture of Experts (SMoE). Les tokens au lieu de tous être traités par un seul modèle vont être répartis selon un réseau de routage vers deux des LLM experts qui pourront traiter leurs tokens, leur sortie étant enfin additionnée pour obtenir un résultat. Mixtral pour ce faire utilise donc 8 modèles 7B comme experts qui sont entraînés en même temps que le réseau de routage. Ce qui permet d'utiliser par traitement un nombre bien plus faible de paramètres (13B) que ce qui est accessible (47B).

IV.4.2 Principaux résultats

MistralAI nous a facilité le travail en publiant une comparaison entre les modèles LLama, Mistral et GPT. Voici ici compilés leurs résultats. Afin d'avoir une comparaison plus neutre, nous avons rajouté en *couleur* les scores publiés par LLama.

		few-Shot	0-Shot	0-Shot	0-Shot	0-Shot	0-Shot	few-Shot	few-Shot	0-Shot	few-Shot	few-Shot	few-Shot
Model	Modality	MMLU	Hella Swag	WinoGPIQA	ARC (easy)	ARC (challenge)	NQ	Trivia QA	Human Eval	MBPP	MATH	GSM8K	
LlaMA 2 7B	Pre trained	44.4	77.1	69.5	77.9	68.7	43.2	24.7	63.8	11.6	26.1	3.9	16.0
		45.3	77.2	69.2	78.8	75.2	45.9	25.7	72.1*	12.8	20.8	2.5	14.6*
LlaMA 2 13B	Pre trained	55.6	80.7	72.9	80.8	75.2	48.8	29.0	69.6	18.9	35.4	6.0	34.3
		54.8	80.7	72.8	80.5	77.3	49.4	31.2	79.6*	18.3	30.6	3.9	28.7*
LlaMA 2 33B	Pre trained	56.8	83.7	76.2	82.2	79.6	54.4	24.1	68.5	25.0	40.9	8.4	44.1
		62.6	83.3	76.7	81.9	79.4	54.5	32.8	84.5*	22.6	33.0	6.24	42.2*
LlaMA 2 70B	Pre trained	69.9	85.4	80.4	82.6	79.9	56.5	25.4	73.0	29.3	49.8	13.8	69.6
		68.9	85.3	80.2	82.8	80.2	57.4	33.0	85.0*	29.9	45.0	13.5	56.8*
Mistral 7B	Pre trained	60.1	81.3	75.3	83.0	80.0	55.5	28.8	69.9	30.5	47.5	13.1	52.2
Mixtral 8x7B	Pre trained	70.6	84.4	77.2	83.6	83.1	59.7	30.6	71.5	40.2	60.7	28.4	74.4
GPT 3 175B	Pre trained	26*	78.1	70.2	80.5 ¹	68.8	51.4	29.9*	71.2	-	-	-	-
GPT 3.5	Pre trained	70.0	85.5*	81.6*	-	-	85.2*	-	-	48.1	-	81.1	57.1
GPT 4	Pre trained	86.4	95.3*	87.5*	-	-	96.3*	-	-	67.0	-	-	92 ¹

¹ Possibilité de contamination des données

*Test réalisé dans des conditions différentes (en few-shot, ou un ensemble de données différente)

00.0:Données tirées des documents de présentation de Mistral et Mixtral

00.0: Données tirées des papiers hors de ceux de Mistral

ce tableau est une combinaison de plusieurs tableaux provenant de diverse documents : Mixtral of Experts, GPT-4 Technical Report, Llama 2 : Open Foundation and Fine-Tuned Chat Models, Mistral 7B, Language Models are Few-Shot Learners

Dans ce tableau regroupant plusieurs informations de différents articles scientifiques. On remarque que Mistral 7B obtient des performances similaires à Llama 2 33B et Mixtral que le modèle 70B, sur l'ensemble des benchmarks réalisés. De plus, en se basant sur les mêmes conditions utilisées par OpenAI, Mistral 7B a également des résultats cotoyant ceux de GPT 3.5.

	Llama 2 70B	Mixtral 8x7B
BBQ accuracy	51.5%	56.0%
BOLD sentiment score (avg ± std)		
gender	0.293 ± 0.073	0.323 ± 0.045
profession	0.218 ± 0.073	0.243 ± 0.087
religious_ideology	0.188 ± 0.133	0.144 ± 0.089
political_ideology	0.149 ± 0.140	0.186 ± 0.146
race	0.232 ± 0.049	0.232 ± 0.052

(cf. Mixtral of Experts)

Concernant le biais dans les réponses données par Mixtral, celles-là ont de meilleurs résultats que

celles fournies par Llama 2. Cependant, cela pourrait être encore amélioré.

Model	Active Params	French			German			Spanish			Italian		
		Arc-c	HellaS	MMLU	Arc-c	HellaS	MMLU	Arc-c	HellaS	MMLU	Arc-c	HellaS	MMLU
LLaMA 1 33B	33B	39.3%	68.1%	49.9%	41.1%	63.3%	48.7%	45.7%	69.8%	52.3%	42.9%	65.4%	49.0%
LLaMA 2 70B	70B	49.9%	72.5%	64.3%	47.3%	68.7%	64.2%	50.5%	74.5%	66.0%	49.4%	70.9%	65.1%
Mixtral 8x7B	13B	58.2%	77.4%	70.9%	54.3%	73.0%	71.5%	55.4%	77.6%	72.5%	52.8%	75.1%	70.9%

(cf. Mixtral of Experts)

Le modèles Mixtral a également été entraîné avec des données multilingues et donc testé sur une collection de références multilingue et on peut voir que les résultats sont très bons également dans d'autres langues en plus de l'Anglais.

IV.4.3 Conclusion

Les modèles proposés par MistralAI montrent une plus grande efficacité que les modèles Llama ou GPT en utilisant moins de paramètre et en réalisant plus rapidement leur traitement. Les biais contenu dans les réponses est plus réduit que les autres modèles et Mixtral et très performant dans plusieurs langues. De plus, ces modèles ont l'avantage d'être sous Licence Apache 2.0 plutôt qu'une licence propriétaire. Tous ces critères sont d'excellentes nouvelles pour notre projet. Si nous partons sur cette piste, il faudra vérifier s'il nous est possible d'utiliser Mixtral ou si nous devons cantonner à Mistral.

V – Conclusion : choix

Suite à nos recherches, nous avons exploré diverses pistes pour la détection automatique de la subjectivité dans les articles de presse. L'architecture Bert-fineTuned, largement adoptée par les meilleures équipes du CheckThat 2023, apparaît comme une solution fiable et aisément implémentable grâce à l'abondance de documentation. Toutefois, étant donné que les performances maximales atteintes par ces équipes ne sont pas spécifiées, il serait judicieux d'envisager d'autres options.

Les architectures combinant BERT avec CNN (surnommé FakeBERT) et BERT avec LSTM montrent un potentiel certain pour cette tâche, en raison de leur aptitude à comprendre des structures de phrases complexes et à identifier des éléments tels que le sarcasme ou la véracité des informations. Ces compétences sont essentielles pour la détection de la subjectivité. Néanmoins, leur efficacité dépend de la disponibilité d'une importante quantité de données d'entraînement, ce qui pourrait limiter leur applicabilité à notre projet, compte tenu de la taille relativement modeste de notre ensemble de données. Une stratégie envisageable consiste à adopter une approche "multilingue" globale, similaire à celle employée par certaines équipes en 2023. La question de développer une architecture neuronale spécifique à nos données demeure ouverte et mérite d'être discutée avec notre client.

Par ailleurs, les modèles de langage à grande échelle (LLM) pré-entraînés, tels que Mistral, ouvrent de nouvelles perspectives. Grâce à leur capacité d'adaptation à des tâches spécifiques avec un minimum de données supplémentaires via le prompt engineering, ces modèles représentent une voie prometteuse à explorer. Bien qu'il n'existe pas encore de documentation sur leur application spécifique à la détection de la subjectivité, leur flexibilité et leur performance en traitement du langage naturel en font des candidats attrayants. Il restera à déterminer le choix entre Mistral et Mixtral.

Au vu de ces éléments, nous suggérons l'ordre de tests suivant :

- Mistral ou Mixtral, pour évaluer leur potentiel innovant dans notre contexte.
- Un modèle FakeBert (BERT+CNN), qui a montré d'excellents résultats dans la détection de fausses nouvelles. Les performances précises (score) restent à spécifier.
- L'architecture Bert-fineTuned, pour bénéficier de résultats éprouvés et fiables.

La priorisation de ces tests sera discutée avec le client. Étant donné l'ampleur du projet et les contraintes de temps, il est possible que nous ne puissions pas explorer chaque piste pour toutes les langues envisagées. Cette démarche stratégique permettra d'aligner nos efforts de recherche et développement avec les objectifs et les ressources disponibles.

VI – Références

1. LLaMA par Meta Ai :
<https://arxiv.org/pdf/2302.13971.pdf>
2. FakeBert :
<https://link.springer.com/article/10.1007/s11042-020-10183-2>
3. Annotation Guidelines for subjectivity detection :
<https://ceur-ws.org/Vol-3370/paper10.pdf>
4. Bert + LSTM :
<https://link.springer-com.gorgone.univ-toulouse.fr/article/10.1007/s10844-022-00755-z>
5. Initiative CLEF :
<https://www.clef-initiative.eu/>
6. CLEF 2023, Task 2 :
<https://checkthat.gitlab.io/clef2023/task2/>
7. Overview CLEF2023 Task 2 :
<https://ceur-ws.org/Vol-3497/paper-020.pdf>
8. GPT-4 Technical Report :
<https://arxiv.org/pdf/2303.08774v4.pdf>
9. Llama 2 : Open Foundation and Fine-Tuned Chat Models : <https://arxiv.org/pdf/2307.09288v2.pdf>
10. Mixtral of Experts :
<https://arxiv.org/pdf/2401.04088v1.pdf>
11. Mistral 7B :
<https://arxiv.org/pdf/2310.06825v1.pdf>
12. Language Models are Few-Shot Learners :
<https://arxiv.org/pdf/2005.14165v4.pdf>