Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240556

Evaluation of BERT-Based Models on Patient Data from French Social Media

Emma LE PRIOL[a,b,c,1], Manissa TALMATKADI[c], Stéphane SCHÜCK[c], Nathalie TEXIER[c] and Anita BURGUN[a,b,d,e]
[a] *HeKA team, Inria, Inserm, France*
[b] *Université Paris-Cité, France*
[c] *Kap Code, Paris, France*
[d] *Clinical Bioinformatics group, Institut Imagine, France*
[e] *Department of Medical Informatics, Necker Hospital, AP-HP, France*
ORCiD ID: Emma Le Priol https://orcid.org/0009-0004-0958-6949, Manissa Talmatkadi https://orcid.org/0000-0002-8019-6159, Stéphane Schück https://orcid.org/0000-0003-2642-7726, Nathalie Texier https://orcid.org/0000-0002-6297-1748, Anita Burgun https://orcid.org/0000-0001-6855-4366

Abstract. With the objective of extracting new knowledge about rare diseases from social media messages, we evaluated three models on a Named Entity Recognition (NER) task, consisting of extracting phenotypes and treatments from social media messages. We trained the three models on a dataset with social media messages about Developmental and Epileptic Encephalopathies and more common diseases. This preliminary study revealed that CamemBERT and CamemBERT-bio exhibit similar performance on social media testimonials, slightly outperforming DrBERT. It also highlighted that their performance was lower on this type of data than on structured health datasets. Limitations, including a narrow focus on NER performance and dataset-specific evaluation, call for further research to fully assess model capabilities on larger and more diverse datasets.

Keywords. transformers, BERT, French, health, word embeddings, rare diseases, developmental and epileptic encephalopathies

1. Introduction

1.1. Background

Social media provide valuable insights about health concerns, living conditions, and clinical histories [1] within general populations or groups sharing a given condition. In Europe, a disease is considered to be rare when affecting fewer than 1 in 2,000 people. Social media have the potential to provide a better understanding of these diseases, including signs, symptoms, and impact on daily life. Among rare diseases, Developmental and Epileptic Encephalopathies (DEEs) are a group of rare and severe epilepsies, associated with frequent seizures of different types, and significant developmental delay, regression or plateau [2].

[1] Corresponding author: Emma Le Priol; E-mail: emma.le-priol@inserm.fr.

Deep learning methods have received attention in recent years for social media analysis and a recent review showed that neural networks perform better than traditional machine learning methods [3]. Large language models exemplified by ChatGPT are undoubtedly powerful tools to discover group themes within text but their performance needs to be evaluated in medical domains as specific as rare diseases and on specific types of data such as social media testimonies. CamemBERT [4], CamemBERT-bio [5] and DrBERT [6] are three language models based on transformers, developed specifically for the French language. CamemBERT-bio, based on CamemBERT, was specifically trained on a French public biomedical dataset [5]; DrBERT is a RoBERTa trained on an open-source corpus of French medical crawled textual data called NACHOS [6].

1.2. Objective

Our aim was to assess the performance of three transformer models, namely CamemBERT [4], DrBERT [6] and CamemBERT-bio [5], on a Named Entity Recognition (NER) task, in which the entities of interest are phenotypes (all types of clinical signs and conditions present in messages posted on social media) and treatments, using social media messages from both patients and caregivers discussing health-related topics. Specifically, the dataset contains testimonies about rare DEEs and messages about common diseases.

2. Methods

2.1. Data and Annotation Process

Two datasets were used and merged.

- A dataset of 16k messages posted between 2013 and 2023 on public forums and social media platforms, extracted using Brandwatch on a list of DEEs based on [7,8]. Among them, five hundred were annotated by one author (ELP) in two steps: (i) regular expressions (regex) were used to detect terms from the Human Phenotype Ontology [9], the UMLS [10] and from a list of drugs; (ii) manual validation and/or addition of terms was made, using PyLighter [11]. During this second step, some messages were excluded because they were not relevant to the field. The resulting training set contained 670 phenotype entities and 67 treatment entities.
- A dataset of 2k messages from patients or caregivers with broader domain coverage (i.e., regardless of any specific disease). Phenotype and treatment entities were manually annotated by eight annotators, with a kappa score of 0.65. In this dataset, there were 3,086 phenotype entities and 1,897 treatment entities.

All the annotations were then adapted to the tokenized texts from the three models, in the IOB2 format. Data was split into a training set (75%), a validation set (15%) and a test set (15%).

2.2. Training

All three models, namely CamemBERT, DrBERT, and CamemBERT-bio were trained using the same architecture. They were trained to optimize the F1-score throughout 10 epochs. The evaluation strategy involved calculating the model's performance after each epoch. The best-performing model, determined by the best F1 on the validation dataset, was kept at the end of the training process. The learning rate was set to 1e-4, controlling the rate at which the model's parameters were updated during training. Both the training and evaluation batches consisted of 8 samples per device. We performed 2020 steps. A weight decay of 0.01 was applied to prevent overfitting during training.

2.3. Evaluation

We averaged the results over 10 runs as follows. For each model and each run, the best model on the validation set was used to calculate the different metrics on the test set. The predictions made on the test set were used to calculate the scores, which were then averaged and presented. Scores were calculated using the seqeval tool on strict mode [12], the same framework used in [5]. In order to compare the models on our NER task, we calculated precision, recall, and F1 for each entity as well as the micro, macro, and weighted averages. Additionally, we compared our micro-average precision, recall, and F1s with those reported in [5] for other NER tasks, namely EMEA - containing drug leaflets - and MEDLINE - containing scientific article titles [13] which were manually annotated following ten semantic groups from UMLS [10].

3. Results

3.1. Comparison of the Models on our Social Media Testimonies NER Task

Table 1. Performance comparison -using seqeval [12] in strict mode- of CamemBERT-base, CamemBERT-bio-base and DrBERT-7GB on NER task on social media testimonies.

		camembert-base	camembert-bio-base	DrBERT-7GB
phenotype	f1-score	0,55	0,56	0,49
	precision	0,62	0,58	0,58
	recall	0,5	0,53	0,43
treatment	f1-score	0,64	0,65	0,64
	precision	0,58	0,56	0,62
	recall	0,72	0,78	0,65
micro-average	f1-score	0,59	0,60	0,55
	precision	0,60	0,57	0,60
	recall	0,58	0,62	0,51
macro-average	f1-score	0,60	0,61	0,57
	precision	0,60	0,57	0,60
	recall	0,61	0,66	0,54
weighted-average	f1-score	0,58	0,59	0,54
	precision	0,61	0,57	0,59
	recall	0,58	0,62	0,51

The evaluation of the three models focused on identifying two categories of entities: phenotypes and treatments in messages posted on social media by patients or caregivers.

Phenotypes. CamemBERT-bio and CamemBERT models demonstrated comparable performance, with F1-scores of .56 and .55 respectively. The three models exhibited a better precision, of .62, .58, and .58, than recall of .50, .53 and .43 respectively. CamemBERT-bio showed the closest recall and precision performance.

Treatments. The three models achieved close F1-scores between .64 and .65 in identifying drug entities. For all models, recall was higher than precision: .65 compared to .62 for DrBERT, .78 compared to .56 for CamemBERT-bio and .72 compared to .58 for CamemBERT.

Micro-average and macro-average. The overall performance of CamemBERT-bio and CamemBERT models was similar, both achieving a micro and macro-average F1-score between .59 and .61. DrBERT achieved slightly lower performance with a micro-average F1-score of .55 and a macro-average F1-score of .57.

Weighted-average. Considering class imbalance, CamemBERT-bio and CamemBERT performed the best, achieving both weighted-average F1-scores of .58-.59.

3.2. Comparison of F1-scores Across NER Tasks

We compared the performance of our task and setting to the ones on EMEA and MEDLINE NER tasks [5]. For all three models, the weighted F1-scores were better for EMEA (.87-.90) and MEDLINE (.76-.78) tasks than for ours (.54-.59). However, they had all three low macro F1-scores on the EMEA (.35-.47) and MEDLINE (.12-.15) tasks, compared to our task (.57-.61).

Table 2. Comparison of F1-scores -using seqeval [12] in strict mode- of CamemBERT-base, CamemBERT-bio-base and DrBERT-7GB on three different NER tasks: EMEA, MEDLINE [5] and our NER task.

	EMEA			MEDLINE			Social Media Testimonies		
	weighted	macro	micro	weighted	macro	micro	weighted	macro	micro
drbert	0,87	0,35	-	0,76	0,15	-	0,54	0,57	0,55
camembert-bio	0,90	0,36	0,77	0,78	0,15	0,68	0,59	0,61	0,60
camembert	0,88	0,47	-	0,76	0,12	-	0,58	0,60	0,59

4. Discussion

Interpretation of Model Performance. The main result of this study is the similar performance achieved by a general model - CamemBERT - and a domain-specific one - CamemBERT-bio - in identifying phenotypes in social media testimonies. These results suggest that both models effectively leverage the language representations learned from French general and health-related data. DrBERT's performance was slightly lower on phenotypes but better on treatments. Overall, the weighted-average F1-score was the same for CamemBERT and CamemBERT-bio and slightly lower for DrBERT.

Comparison Across NER Tasks in Different Contexts. The comparison of F1-scores across three different NER tasks - EMEA, MEDLINE, and our social media testimonies task - indicates that the weighted F1-scores for both EMEA and MEDLINE tasks were notably higher (.87-.90 and .76-.78, respectively) compared to our social media testimonies task (.54-.59). This discrepancy suggests that the models perform better

when applied to more structured and domain-specific datasets, such as those found in regulatory or biomedical literature contexts.

5. Conclusions

This preliminary work revealed that on social media testimonies, CamemBERT and CamemBERT-bio have almost the same performance, slightly higher than DrBERT. It also highlighted that the performance on social media testimonies is still lower than on structured and more traditional health datasets like MEDLINE. Models capable of accurately extracting phenotypes and treatment entities from social media messages could help in better understanding the challenges faced by patients and their caregivers, especially in the rare disease and DEEs field.

Several limitations should nevertheless be addressed. The evaluation primarily focused on NER performance, neglecting other important tasks. Future research is required to provide a more comprehensive assessment of model capabilities. Additionally, the generalizability of our findings may be limited by the specific dataset and evaluation setup used in this study. Further research involving larger and more diverse datasets is needed to validate and extend our findings.

Acknowledgment: This work was supported by state funding as part of the "Investissements d'avenir" program (ANR-19-P3IA-0001) (PRAIRIE 3IA Institute).

References

[1] Klein AZ, Gutiérrez Gómez JA, Levine LD, Gonzalez-Hernandez G. Using Longitudinal Twitter Data for Digital Epidemiology of Childhood Health Outcomes: An Annotated Data Set and Deep Neural Network Classifiers. J Med Internet Res 2024;26:e50652. doi:10.2196/50652.
[2] Developmental Epileptic Encephalopathy. n.d. https://epilepsyfoundation.org.au (accessed April 2, 2024).
[3] Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. Npj Digit Med 2022;5:46. doi:10.1038/s41746-022-00589-7.
[4] Martin L, Muller B, Suárez PJO, Dupont Y, Romary L, de la Clergerie ÉV, et al. CamemBERT: a Tasty French Language Model. Proc. 58th Annu. Meet. Assoc. Comput. Linguist., 2020, p. 7203–19. doi:10.18653/v1/2020.acl-main.645.
[5] Touchent R, Romary L, de la Clergerie E. CamemBERT-bio: a Tasty French Language Model Better for your Health 2023. doi:10.48550/arXiv.2306.15550.
[6] Labrak Y, Bazoge A, Dufour R, Rouvier M, Morin E, Daille B, et al. DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains 2023. doi:10.48550/arXiv.2304.00958.
[7] Gallop K, Lloyd AJ, Olt J, Marshall J. Impact of developmental and epileptic encephalopathies on caregivers: A literature review. Epilepsy Behav 2021;124:108324. doi:10.1016/j.yebeh.2021.108324.
[8] Salom R, et al. Dataset on the psychosocial impact in families with children with developmental and epileptic encephalopathies. Sci Data 2023;10:530. doi:10.1038/s41597-023-02441-3.
[9] Gargano MA, Matentzoglu N, et al. The Human Phenotype Ontology in 2024: phenotypes around the world. Nucleic Acids Res 2024;52:D1333–46. doi:10.1093/nar/gkad1005.
[10] Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Yearb Med Inform 1993:41–51. doi:10.1055/s-0038-1637976.
[11] pylighter: Annotation tool for NER tasks on Jupyter n.d.
[12] seqeval: Testing framework for sequence labeling n.d.
[13] Névéol A, Grouin C, Leixa J, Rosset S, Zweigenbaum P. The Quaero French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization n.d.