IVAMED: Intelligent Virtual Assistant for Medical Diagnosis

Dana Gallent-Iglesias[1], Santiago Serantes-Raposo[1], Iñigo López-Riobóo-Botana[1], Sonia Gonzalez-Vázquez[1] and Pablo Manuel Fernandez-Graña[1]

[1] *Instituto Tecnológico de Galicia - ITG - Centro Tecnológico Nacional, Cantón Grande 9, Planta 2, 15003, A Coruña, España*

Abstract

The recent advancements in deep learning have led to a myriad of approaches for medical diagnosis and assistance. Some topics such as data labelling, data curation, human-in-the-loop, explainability or privacy-preserving methodologies are hot topics for applied machine learning in the healthcare context. In this domain, we normally expect a three-way interaction (doctor-system-patient), so that interfaces play a crucial role. Remotely managed VR (Virtual Reality) systems help us to enhance communication and feedback between doctors and patients in situations where in-person assistance is not feasible. Moreover, the recent breakthroughs with LLMs (Large Language Models) enable us to use natural language as additional interface, considering NLU (Natural Language Understanding) for intent recognition, ASR (Automatic Speech Recognition) and TTS (Text-To-Speech).

In the context of the CEL.IA network, we present **IVAMED** (**I**ntelligent **V**irtual **A**ssistant for **ME**dical **D**iagnosis), a chatbot-oriented application in a VR environment for remote medical assistance. We tackle the situation in which face-to-face assistance is not possible. We provide the tools for remote interaction and guided diagnosis. We propose the evaluation of the MoCA (Montreal Cognitive Assessment) test for early detection of MCI (Mild Cognitive Impairment) and the BDI (Beck Depression Inventory) test for measuring characteristic attitudes and symptoms of depression.

Keywords

IVA, VR, chatbot, Rasa, medical diagnosis, healthcare, BDI test, MoCA test, NLU, ASR, TTS

1. Introduction

In recent years, there has been a significant rise in the use of chatbots in the healthcare industry [1, 2, 3, 4]. Healthcare chatbots offer a convenient and efficient way to interact with patients, providing them with personalised and immediate attention. These chatbots can be integrated into websites or mobile applications, providing round-the-clock assistance to patients. Additionally, the recent advancements in NLP (Natural Language Processing) have led to the development of LLMs (Large Language Models), which can understand the nuances of human language and generate responses that are contextually appropriate.

Healthcare chatbots [5, 6] can be used for tasks such as symptom assessment or medication reminders [7, 8]. They can also provide information on healthcare procedures and preventive measures, which can empower patients to take better care of themselves [9, 10]. Chatbots can also assist healthcare professionals by automating administrative tasks, such as scheduling appointments and managing patient records, allowing them to focus on providing quality care. The healthcare industry has seen significant benefits from chatbots, including reduced wait times, improved patient outcomes, and reduced costs.

There are several standardised cognitive assessment tools used in healthcare, such as the MoCA (Montreal Cognitive Assessment) [11] test and the BDI (Beck Depression Inventory) [12] test. These tests are used to evaluate cognitive and emotional functions, which can aid in diagnosing and treating various conditions. MoCA is a widely used cognitive screening tool that assesses various cognitive domains, such as attention, memory, language, visuospatial abilities, and executive functions. In contrast, BDI is a self-reported questionnaire used to measure the severity of depression symptoms.

VR (Virtual Reality) has been used in healthcare to improve patient outcomes [13], including pain management and rehabilitation [14, 15]. VR can provide a realistic and immersive environment for patients to distract them from their pain and facilitate relaxation. It can also provide a safe and controlled environment for rehabilitation exercises. Chatbots can be integrated with VR technology to provide personalized assistance during their VR sessions.

In Section 1.1, we present our motivation to carry out this project. In Section 1.2 we enumerate our main contributions. In Section 2, we depict the architecture and methodology followed for this project, describing our

SEPLN-PD 2023: Annual Conference of the Spanish Association for Natural Language Processing 2023: Projects and System Demonstrations

dgallent@itg.es (D. Gallent-Iglesias); sserantes@itg.es (S. Serantes-Raposo); ibotana@itg.es (I. López-Riobóo-Botana); sgonzalez@itg.es (S. Gonzalez-Vázquez); pgrana@itg.es (P. M. Fernandez-Graña)
https://www.linkedin.com/in/dana-gallent-iglesias-7a0038247/ (D. Gallent-Iglesias); https://www.linkedin.com/in/%C3%AD%C3%B1igo-luis-l%C3%B3pez-riob%C3%B3o-botana-4a43001a2/ (I. López-Riobóo-Botana); https://www.linkedin.com/in/phd-sonia-gonz%C3%A1lez-v%C3%A1zquez-38b14a8b/ (S. Gonzalez-Vázquez)
0000-0002-7310-0702 (I. López-Riobóo-Botana); 0009-0006-2782-2567 (S. Gonzalez-Vázquez)

results in Section 3. Finally, we conclude with some limitations in Section 4 and future work in Section 5.

1.1. Motivation

For domain-specific sensitive contexts like healthcare, we need to adjust and control the chatbot output thoroughly. The chatbot-oriented LLMs fit in AGI (Artificial General Intelligence) contexts and they can be somewhat "fine-tuned" with prompt engineering [16, 17], but this is not enough for guided and ad hoc ADI (Artificial Domain Intelligence) systems. Another important point of our project is the contribution in the development of remote medical diagnosis tools with VR integration. This enables us to enhance communication and feedback between doctors and patients in situations where in-person assistance is not feasible.

In the context of the CEL.IA network[1], we propose an IVA (Intelligent Virtual Assistant) integrated in a VR environment for remote medical diagnosis. We mixed the concepts of NLU using the Rasa chatbot framework with both ASR and TTS modules. All these components are part of the VR system, which enables the users to perform the BDI and the MoCA tests. In this paper, we describe our first demo version, including the dialogue system in a chatbot-oriented application.

1.2. Contributions

For this IVA project, our contributions are as follows:

- **We integrated a chatbot-oriented application for remote medical diagnosis in a VR environment**, providing visual interaction with selection and manipulation strategies to perform the medical tests (MoCA and BDI).
- **We implemented a domain-specific IVA system with Rasa framework** for NLU and dialogue management.
- **We provide both voice and text interfaces to better communicate with the chatbot**, so that we can use spoken language in Spanish (ASR), as well as listen to the answers (TTS).

2. Methodology

We designed and implemented the following modules:

- **NLU subsystem:** In charge of the intent recognition NLP task. Rasa projects follow a data-driven approach, providing several files with text samples for each intent and configuration files to adjust the pipelines for model training.
- **Dialogue module:** Relies on a combination of a rule-based system and a "user stories" mechanism to infer the next action in the conversation. These actions can be (1) direct chatbot responses or (2) delegations in the SDK action server.
- **ASR module**: We have integrated a speech recognition module, which allows us to convert from audio to text and transmit it back to the chatbot. The text is then analysed by the NLU module and a response is generated by the Rasa dialogue system. We leverage the Spanish *stt_es_citrinet_512* model from the NVIDIA NeMo toolkit[2].
- **TTS module**: After receiving the response from the chatbot, we convert the text output back to human voice, making it convenient to engage with the IVA system. We made use of the Spanish *glow-speak:es_tux* model from the OpenTTS framework[3].

The general diagram of the IVA system is depicted in Figure 1. In Section 2.1, we study in more detail the NLU training pipeline for the chatbot and in Section 2.2, we summarise the VR system integration.

2.1. NLU pipeline

We used some of the Rasa components to train the model [18]. We configured different pipelines, **replacing the feature extractor component and preserving intact the rest of them**:

1. **Tokenizer**: Rasa component in charge of splitting each sentence into tokens or words. We used the simple *WhiteSpaceTokenizer* to get the tokens splitting using white spaces.
2. **Feature extractor**: Rasa component in charge of feature engineering to transform the corresponding tokens into numerical vector representations. We made use of several feature extractors. We provide all the details in the experimental Section 3.
3. **Intent classifier**: Rasa component in charge of the intent recognition NLP task. We used the DIET classifier from the Rasa framework authors [19]. DIET is a multi-task modular transformer architecture that handles both intent classification and entity recognition together. It provides the ability to plug and play various pre-trained embeddings like BERT (and variants), GloVe, ConveRT, among others [20].
4. **Fallback classifier**: This Rasa component is in charge of triggering the default intent when the

[1] https://www.redcelia.es/

[2] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_es_citrinet_512

[3] https://github.com/synesthesiam/opentts#voices

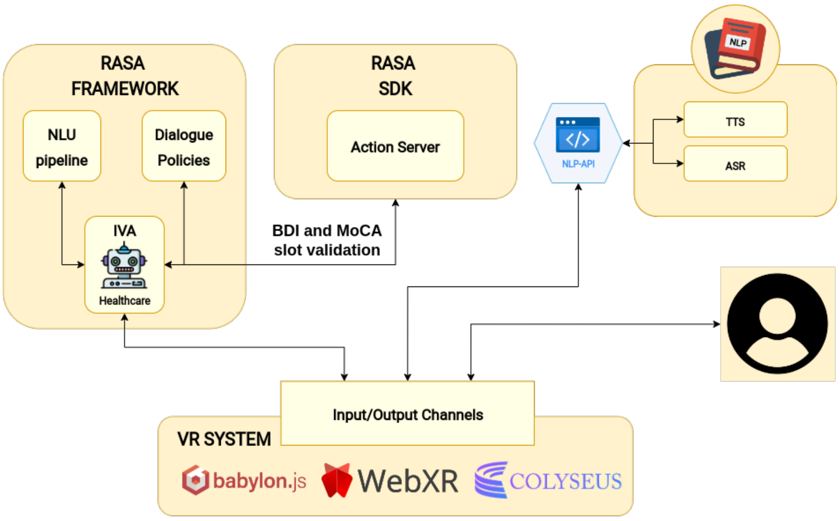

Figure 1: Architecture of the IVA system, **including the integration with the VR system. The Rasa framework is in change of the NLU (intent recognition) task**. The dialogue subsystem is responsible for providing each answer and for guiding the user towards the next input. The **Rasa SDK server automatically validates test slots by means of custom actions**, so that we can generate a final result when the user finishes one of the tests. **The communication interfaces include the VR system and the input/output channels**, using not only text but also voice (ASR and TTS services).

intent prediction from the DIET classifier does not have a confidence above a pre-specified threshold. If so, the default intent will diverge from the regular conversation, following a custom action for managing the unexpected input.

Nowadays, the most amount of NLP approaches **are end-to-end and avoid the feature extraction step [21]**. However, we are following a more classical approach since we do not have enough data for an end-to-end fine-tuning process with transformer models. Moreover, relying on multilingual deep word embeddings as the input features for the DIET classifier has some advantages, considering that this model outperforms fine-tuned BERT and is about six times faster [20].

2.2. Immersive interface

We can highlight the following two main features:

1. **VR environment for medical diagnosis and cognitive tests**: The interface for the cognitive tests (MoCA and BDI tests), as well as the other generic UI (User Interface) elements, make full use of the *Babylon.js* framework, which provides support for the WebXR standard to develop immersive web experiences, deployable on dedicated immersive hardware or web standard. The majority of the UI elements have been implemented as virtual objects within a 3D environment running inside an HTML canvas on the web page. This approach facilitates the switch between a traditional 3D web application with WebGL and an immersive environment with WebXR. Additionally, the scene includes the avatar for the IVA system with a recording button through which audio can be sent to communicate with the conversational IVA and answer to the cognitive test questions (see Figure 2).
2. **Remote assistance service**: The diagnosis system integrates a service that the healthcare professionals can remotely control to evaluate the cognitive tests the patient is carrying out. This service has been built on the Colyseus framework[4], an open source project that provides tools for the implementation of multi-user online experiences. This functionality allows the healthcare professional to create a shared online environment, where a 3D scene of a virtual clinic is loaded. From this virtual space, the doctor has access to a menu where he/she can select the cognitive test to be loaded on the scene and to be performed by the patient. Then, the professional can use this virtual space to visualise patient's progress within each test in real-time. For instance, when the patient is asked to draw a clock, the doctor can observe the drawing's progression in real-time (see Figure 3).

[4] https://www.colyseus.io/colyseus

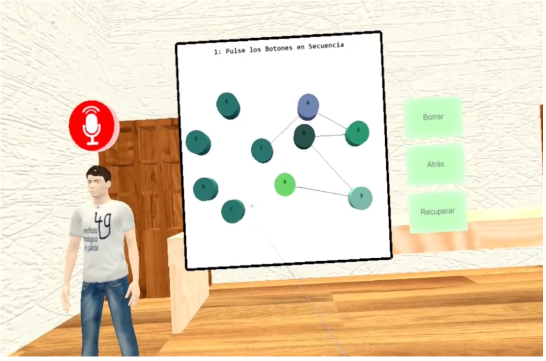

Figure 2: Immersive interface for medical diagnosis.

3. Results

We trained a text classification model for intent recognition to carry out the BDI and MoCA cognitive tests, and so we did for assistance generating a medical report, the three functionalities included in the health demonstrator.

We conducted tests with different configurations for feature extraction in the NLU pipeline, following two different approaches:

1. **Traditional feature extractors**: We made use of (1) **lexical/semantic-based** feature extractors to generate shallow feature vectors, (2) **regex and pattern-based** feature extractors to generate binary vectors according to the pattern being present or not and (3) **Bag-of-words vectors** with 1-grams words and 4-grams characters [22].
2. **Pre-trained language models feature extractors**: Based on the transformer architecture for obtaining deep word embeddings. We considered the models BERT (Spanish monolingual version) [23], DistilBERT (multilingual version) [24], XML-RoBERTa (multilingual version) [25] and GPT-2 (Spanish monolingual version) [26].

For the intent classification task, we used the DIET classifier [20]. We compared the results obtained following the two aforementioned feature extraction approaches, defining 5 different configurations: BERT, DistilBERT, GPT-2 and XLM-RoBERTa for the pre-trained language models feature extractors and the "No model" approximation for the traditional feature extractors. We followed a 4-fold cross-validation. We computed the macro average for the f1-score, which summarises the recall and precision metrics of the intent classifier.

Intuitively, feature extraction using transformers and their deep word embeddings should be better. However, looking at the Figure 4, this is not the case. The highest macro f1-score value was achieved by the "No model" approach, which uses the traditional feature extractors. The classic feature extractors behave better with short and simple sentences, where the global context is not relevant and the vocabulary is limited (MoCA and BDI tests for medical diagnosis). Among the transformer-based extractors, the worst results were achieved by GPT-2, since this is a decoder-only transformer model optimised for NLG (Natural Language Generation) tasks rather than for NLU classification tasks. It is worth mentioning that the best transformer approach was DistilBERT, the light-weight encoder-only model needing less computing resources, even outperforming the XLM-RoBERTa model for feature extraction.

4. Limitations

Some problems could arise when using our chatbot implementation:

- **Rasa Knowledge base**: Due to the data-driven nature of Rasa, any new example must be manually added to the knowledge base and, consequently, it is required to retrain a new model including the changes in the NLU pipeline.
- **Traditional feature extraction vs deep word embeddings**: As it was exposed in the Section 3, the simplicity of the sentences expected by our chatbot favours traditional feature extractors over transformers and deep embeddings. Using a language model limits the chatbot performance if we are not considering an end-to-end approach with a fine-tuned model for the NLU task [21].
- **Similar intent classification**: Although this limitation was mitigated implementing similarity scores (in order to match each user input to the most reasonable answer for each medical test slot), the chatbot still occasionally struggles with classifying intents when the sentence is almost identical to an example from another intent.

5. Conclusion and future work

In this work, we present **IVAMED**, a chatbot-oriented IVA application for medical diagnosis in a VR environment with ASR, TTS and web interfaces. We included a dialogue system capable of guiding the user to perform each of the medical tests and get a result. We made use of the Rasa framework and proposed several NLU pipelines for intent classification.

In the near future, we want to explore some NLU improvements following an end-to-end approach with a fine-tuned intent classifier based on transformer-encoders. In this way, we will avoid the current feature

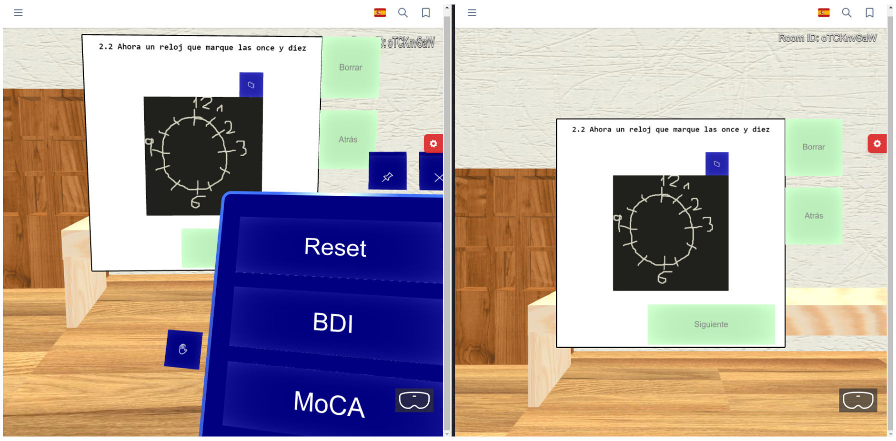
2.2 Ahora un reloj que marque las once y diez
Borrar
Atrás
Reset
BDI
MoCA
2.2 Ahora un reloj que marque las once y diez
Room ID: oTCKnvSaW
Borrar
Atrás
Siguiente

Figure 3: Immersive interface for the clock-drawing test in MoCA [11]. On the left, the doctor's view with a menu panel to select the cognitive test to be loaded. On the right, the view of the patient, performing the test. Notice that the doctor has a real-time image of the drawing the patient is making.

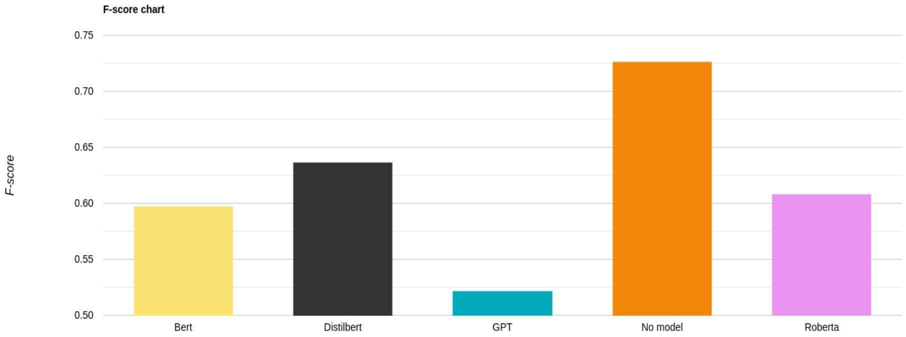

Figure 4: BERT, DistilBERT, GPT-2, "No model" and XLM-RoBERTa pipeline configurations. Since we considered all intents equally important, we present the macro average for the f1-score, following a 4-fold cross validation.

extraction step in the NLU pipelines. We are also exploring data augmentation techniques to increase the number of samples for each intent, leveraging LLMs and cross-language translation models.

Acknowledgments

This project belongs to the CEL.IA network initiative[5], which is supported by the Ministerio de Ciencia e Innovación through the CDTI (Centro para el Desarrollo Tecnológico Industrial) (grant CER-20211022).

References

[1] A. S. Pillai, P. S. Mathew, Impact of virtual reality in healthcare: a review, Virtual and augmented reality in mental health treatment (2019) 17–31.

[2] H. A. Aziz, Virtual reality programs applications in healthcare, Journal of Health & Medical Informatics 9 (2018) 305.

[3] P. Kantithammakorn, P. Punyabukkana, P. N. Pratanwanich, S. Hemrungrojn, et al., Using automatic speech recognition to assess Thai speech language fluency in the Montreal cognitive assessment (MoCA), Sensors 22 (2022) 1583.

[4] C. Fertleman, P. Aubugeau-Williams, C. Sher, A.-N.

[5] https://itg.es/cervera-celia/

Lim, et al., A discussion of virtual reality as a new tool for training healthcare professionals, Frontiers in public health 6 (2018) 44.

[5] N. Bhirud, S. Tataale, S. Randive, S. Nahar, A literature review on chatbots in healthcare domain, International journal of scientific & technology research 8 (2019) 225–231.

[6] S. Laumer, C. Maier, F. T. Gubler, Chatbot Acceptance in Healthcare: Explaining User Adoption of Conversational Agents for disease Diagnosis, in: J. vom Brocke, S. Gregor, O. Müller (Eds.), 27th European Conference on Information Systems - Information Systems for a Sharing Society, ECIS 2019, Stockholm and Uppsala, Sweden, June 8-14, 2019, 2019. URL: https://aisel.aisnet.org/ecis2019_rp/88.

[7] T. Nadarzynski, O. Miles, A. Cowie, D. Ridge, Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study, Digital health 5 (2019) 2055207619871808.

[8] L. Athota, V. K. Shukla, N. Pandey, A. Rana, Chatbot for Healthcare System Using Artificial Intelligence, in: 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020, pp. 619–622. doi:10.1109/ICRITO48877.2020.9197833.

[9] K.-J. Oh, D. Lee, B. Ko, H.-J. Choi, A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation, in: 2017 18th IEEE International Conference on Mobile Data Management (MDM), 2017, pp. 371–375. doi:10.1109/MDM.2017.64.

[10] K. Chung, R. C. Park, Chatbot-based heathcare service with a knowledge base for cloud computing, Cluster Computing 22 (2019) 1925–1937.

[11] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, et al., The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment, Journal of the American Geriatrics Society 53 (2005) 695–699.

[12] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An inventory for measuring depression, Archives of general psychiatry 4 (1961) 561–571.

[13] D. Freeman, S. Reeve, A. Robinson, A. Ehlers, et al., Virtual reality in the assessment, understanding, and treatment of mental health disorders, Psychological medicine 47 (2017) 2393–2400.

[14] G. Tao, B. Garrett, T. Taverner, E. Cordingley, C. Sun, Immersive virtual reality health games: a narrative review of game design, Journal of NeuroEngineering and Rehabilitation 18 (2021) 1–21.

[15] M. Ma, H. Zheng, Virtual reality and serious games in healthcare, Advanced computational intelligence paradigms in healthcare 6. Virtual reality in psychotherapy, rehabilitation, and assessment (2011) 169–192.

[16] S. Qiao, Y. Ou, N. Zhang, X. Chen, et al., Reasoning with Language Model Prompting: A Survey, 2022. URL: https://arxiv.org/abs/2212.09597. doi:10.48550/ARXIV.2212.09597.

[17] Q. Dong, L. Li, D. Dai, C. Zheng, et al., A Survey on In-context Learning, 2023. URL: https://arxiv.org/abs/2301.00234. doi:10.48550/ARXIV.2301.00234.

[18] Rasa, Rasa Components, 2023. URL: https://rasa.com/docs/rasa/components/.

[19] T. Bunk, D. Varshneya, V. Vlasov, A. Nichol, DIET: Lightweight Language Understanding for Dialogue Systems, 2020. URL: https://arxiv.org/abs/2004.09936. doi:10.48550/ARXIV.2004.09936.

[20] M. Mantha, Introducing DIET: state-of-the-art architecture that outperforms fine-tuning BERT and is 6X faster to train, 2020. URL: https://rasa.com/blog/introducing-dual-intent-and-entity-transformer-diet-state-of-the-art-performance-on-a-lightweight-architecture/.

[21] A. Rahali, M. A. Akhloufi, End-to-End Transformer-Based Models in Textual-Based NLP, AI 4 (2023) 54–110. URL: https://www.mdpi.com/2673-2688/4/1/4. doi:10.3390/ai4010004.

[22] Y. Zhang, R. Jin, Z.-H. Zhou, Understanding bag-of-words model: a statistical framework, International journal of machine learning and cybernetics 1 (2010) 43–52.

[23] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[24] HuggingFace, How to use DistilBERT, 2023. URL: https://github.com/huggingface/transformers/tree/main/examples/research_projects/distillation#how-to-use-distilbert.

[25] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, et al., Unsupervised Cross-lingual Representation Learning at Scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners (2019).