EMPHASIS: Empowering Decision Making with Higher Productivity by Means of HyperAutomation

Montse Cuadros[1], Aitor Álvarez[1], Naiara Pérez[2], Juan Manuel Martín[1], Pablo Turón[1], Elena Zotova[1], Haritz Arzelus[1], Joaquin Arellano[1] and Arantza del Pozo[1]

[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi Pasealekua 57, Donostia/San-Sebastian, 20009, Spain
[2]HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Manuel Lardizabal Pasealekua 1, Donostia/San-Sebastian, 20018, Spain

Abstract

The EMPHASIS project aims to support companies in the hyper-automation of tasks by researching and developing tools adaptable to different use cases using speech and natural language processing technologies. The toolset is deployed on the EMPHASIS platform, allowing the building of fully customisable and flexible processing pipelines with a stack of different text and speech components adaptable to specific domains and languages. All the adaptation of components is available as a feature in the main EMPHASIS tools, allowing end-users to adjust the technology to the domains and languages required by each application scenario.

Keywords

NLP, NER, Text classification, Document classification, Question Answering, Hyperautomatization, ASR, Diarization, Emotions

1. Introduction

Since the first industrial revolution, we have been undergoing a continuous process of automating tasks through the application of technology. Today, it is indisputable that the key to the present and the future is software. The main achievements, challenges and opportunities currently focus on the application of digital technologies of different levels of maturity, complexity and supported on different architectures and communication networks. In the strategic plans of the industrial sectors with the greatest weight in GDP (e.g., public administration, and key services like Education and Health), digital transformation is a constant, and it always has a very relevant and growing weight alongside the automation of processes.

However, in the back-office, organisations continue to be intensive in repetitive office work, such as processing documents manually and answering calls (still far ahead of services based on chatbots). The reason is simple: RPA [1] systems can automate routine and repetitive tasks, but they have a lot of problems when it comes to making decisions, providing personalised non-template-based responses, etc. This is largely due to the fact that the information that is still handled in the back-office today consists of documents and/or calls, i.e., unstructured information on which the most basic automation systems find it difficult to act and make decisions. Therefore, the back-office still relies heavily on human effort and knowledge, which prevents us from moving towards the data economy. This is where Artificial Intelligence (AI), and specifically Natural Language Processing (NLP), can play a key role as a complement to RPA systems, in what has come to be called Hyperautomation [2] or Cognitive Automation, which is the integration of speech and language processing, AI, and cognitive services with RPA systems. It is about carrying out a total digital transformation, implementing in the office, in the back-office, processes that scale in the same way as the rest of the processes of the systems (customer, plant, etc.). In this context, we present the EMPHASIS project (from 'EMpowering decision making with higher productivity by means of hyper-Automation-based SolutIonS'). EMPHASIS has developed and researched speech and NLP technologies adapted to real use cases, where they are integrated in a platform easy to use and adaptable to use cases.

This paper is organised as follows, section 2 presents the consortium and funding body, section 3 gives an overview of the goals and expected results, section 4 draws the main challenges. Section 5 gives an overview of the main use cases tackled in the project and section 7 explains the main technological components developed. Finally, section 8 highlights the main conclusions and future work.

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain.

mcuadros@vicomtech.org (M. Cuadros); aalvarez@vicomtech.org (A. Álvarez); naiara.perez@ehu.eus (N. Pérez); jmmartin@vicomtech.org (J. M. Martín); pturon@vicomtech.org (P. Turón); ezotova@vicomtech.org (E. Zotova); harzelus@vicomtech.org (H. Arzelus); jarellano@vicomtech.org (J. Arellano); adelpozo@vicomtech.org (A. d. Pozo)

0000-0002-3620-1053 (M. Cuadros); 0000-0002-7938-4486 (A. Álvarez); 0000-0001-8648-0428 (N. Pérez); 0000-0003-4874-0166 (J. M. Martín); 0000-0002-5563-1120 (P. Turón); 0000-0002-8350-1331 (E. Zotova); 0000-0002-0731-1317 (H. Arzelus); 0000-0003-3505-5514 (A. d. Pozo)

2. Consortium and funding body

EMPHASIS has been partially funded by the Basque government through the Hazitek Estratégico 2021 programme of the SPRI Group under grant agreement ZE-2021/00039. It has run from 05/2021 to 12/2023. The consortium is led by Grupo Teknei[1], Ibermatica[2], Natural Vox[3], Eutik[4], Gureak[5], Segula[6], Zuchetti[7] and Merkatu[8]. The research centres involved are Vicomtech, Ibermatica's R&D business unit (I3B) and the Speech Interactive Group from the UPV/EHU[9].

3. Goals and expected results

The aim of this project is to take the back-office towards what has been called Hyperautomation or Cognitive Automation, which proposes to integrate AI with RPA solutions.

The integration of speech and natural language processing technologies makes it possible to structure content that has not been structured until now, such as documents (invoices, delivery notes, deeds, building permits, municipal licences, etc.) without having to have an associated template, or calls (in call centres, for example). It is a matter of converting all this content (documents and calls) into data, so that it can be managed, analysed and the associated processes can be optimised, in order to make progress in a real digital transformation, which otherwise seems impossible. It is not a matter of doing digitally the same as we do now, but rather, based on data, defining the optimal flows and processes that guarantee total digitisation, scalable, robust, flexible in the face of changes, adaptable to new scenarios, and always with reasonable costs. To this end, another key factor is to follow the principles of no code or, failing that, low code, i.e., to develop solutions whose integration and implementation are as transparent as possible, avoiding the costly processes involved in BPM (Business Process Management) solutions.

The aim is therefore to develop technology that can be integrated as APIs in third-party solutions, so that their configuration can be carried out by non experts. In the same way, the aim is for these APIs to encapsulate functions that can be customised, i.e., one of the main objectives of the project is that the functions developed within the framework of the project can later be customised and adapted to different use cases, following the principles of low-code.

4. Challenges

The main technological challenges of EMPHASIS are related to the application of the following technologies to different languages, domains, audio and text formats, together with the development of tools to allow their easy adaption to the real use case of each client:

- **Cognitive Document Automation:** its objective is to automate the extraction of relevant content from textual sources of different formats (invoices, delivery notes, e-mails, text documents, FAQ systems, conversations extracted by automatic transcription [3]), to classify content into different categories [4], domains, feelings, and also extract core-entities (NER) and their relationships that allow the understanding of its content. The technology to be used is cutting-edge Deep Learning technology where a paradigm shift has been seen in the last two years thanks to the proliferation of advanced neural architectures that exploit language models with knowledge of the world. Documents containing images with text inside are also converted, and their reading is improved by applying advanced methods on state-of-the-art OCR tools such as LayoutLM [5]. Additionally, technology based on Question Answering techniques can also be used in order to have tools to search into documents and find relevant information.
- **Speech Analytics:** its objective is to automate the analysis of data with acoustic content by transformation to text using advanced tools for automatic enriched speech transcription and emotion analysis [6]. The main technology to be used is Deep Learning technology based on the latest contributions from the scientific community to the state of the art and which has shown great improvements with respect to previous Machine Learning technology.

5. Use cases

EMPHASIS proposes a set of application scenarios that form a common denominator in terms of technological needs and will allow us to join forces towards the development of a global solution for the cognitive automation of processes related to documents and audios in Spanish. The main application scenarios can be divided at a more technological level into two subgroups where:

[1] https://www.teknei.com/
[2] https://ibermatica.com/
[3] https://naturalspeech.es/
[4] https://www.eutik.com/
[5] https://www.gureakmarketing.com/es/
[6] https://www.segulatechnologies.com/es/
[7] https://www.zucchetti.es/
[8] https://www.merkatu.com/
[9] https://www.ehu.eus/en/web/speech-interactive

- Text processing technology is linked to Cognitive Document Automation (CDA).
- Speech processing technology is fundamental to carry out cognitive automation related to audios, phone calls, etc. also called Speech Analytics.

All in all, these technologies have been divided into different verticals or domains related to use-cases that have been defined in EMPHASIS by the different companies participating in the project:

Banking: Speech Analytics is used to work on solutions that allow the acoustic analysis of audios through the enriched transcription of their content for subsequent classification of customers in quadrants. In this use case, the transcription of the content will automate the management processes of customer requests and their correct attention. In the case of Text Analyics, the detailed analysis of textual documents to make a correct segmentation of their content, classify them by language and typology and extract the metadata. In addition, there is a need to go a step further and manage documents in order to automate the search for documents related to questions, in environments such as FAQs.

Retail: The main objective of the project has been to provide technology for the automatic processing of documents such as invoices, delivery notes to facilitate the integration of the content of these documents into an ERP automatically or with minimum supervision.

Justice: Using Speech Analytics, the project aims to analyse audios in the legal field, through enriched transcription, identification of speakers, through biometrics [7], detection of emotions in the audios. In addition to creating transcription routines with time stamps to quickly locate the parts corresponding to a transcribed text. In terms of text analysis, there is a huge need of using Anonymization [8].

Administration: The main objectives are the detailed analysis of textual documents to make a correct segmentation of their content, classify them by language and typology and extract the metadata. In addition, there is a need to go a step further and manage documents in order to automate the search for documents related to questions, in environments such as FAQs.

Industry: In relation to the use of Speech Analytics, the main challenge is to classify incidents by voice and in real time. In terms of Textual Analytics, the main challenges to be worked on in the project are the classification of textual documents, analysis of e-mails and attachments to manage content and classify it according to the specifications of each domain.

Contact Centres: In relation to the use of Speech Analytics, the main challenges are the analysis of customer responses to VoiceBots to improve the processes of automatic interpretation of responses and to analyse their emotions. In addition, if the interaction is agent-based, early detection of satisfaction levels or performance compliance. Other channels to analyse apart from audio are textual channels, including chats, e-mails or social networks, which are sources of data that will be used to analyse user satisfaction and decision-making to improve the solution's knowledge flows in this use case or to interact with the customer automatically in search of information. In this sense, texts will be analysed in order to classify them by subject matter. Finally, the last channel to be used in this use case is the physical document, which must be converted into a textual document using computer vision and OCR techniques.

6. Approach

Therefore, we are talking about developing basic solutions for document and audio processing, but at the same time, we are talking about developing tools and systems that allow the customisation of these solutions. To give an example, if one of the base solutions to be generated allows the classification of documents and whether it is an invoice or a delivery note, tools will also be generated to help in the process of annotating other documents and generating the corresponding AI models, so that in the future the customised solution can classify documents by identifying whether it is a building permit, a deed, or a municipal licence. All modules developed in the project have been deployed as REST APIs within a Docker container, so each module could be used as a single tool. However, the project has developed a core platform where different sets of tools can be concatenated as fully configurable pipelines. This allows building pipelines per domain and per language based on each partner or each use-case needs. Figure 1 shows a diagram of the EMPHASIS platform. Last but not least, the platform allows resource, queue, and pipeline management, and also load-balance configurations.

7. Main technological modules

In this section, we present the main technological modules developed in the framework of the EMPHASIS project based on Text and Speech Technologies, which are integrated, as explained in section 6, into the EMPHASIS platform.

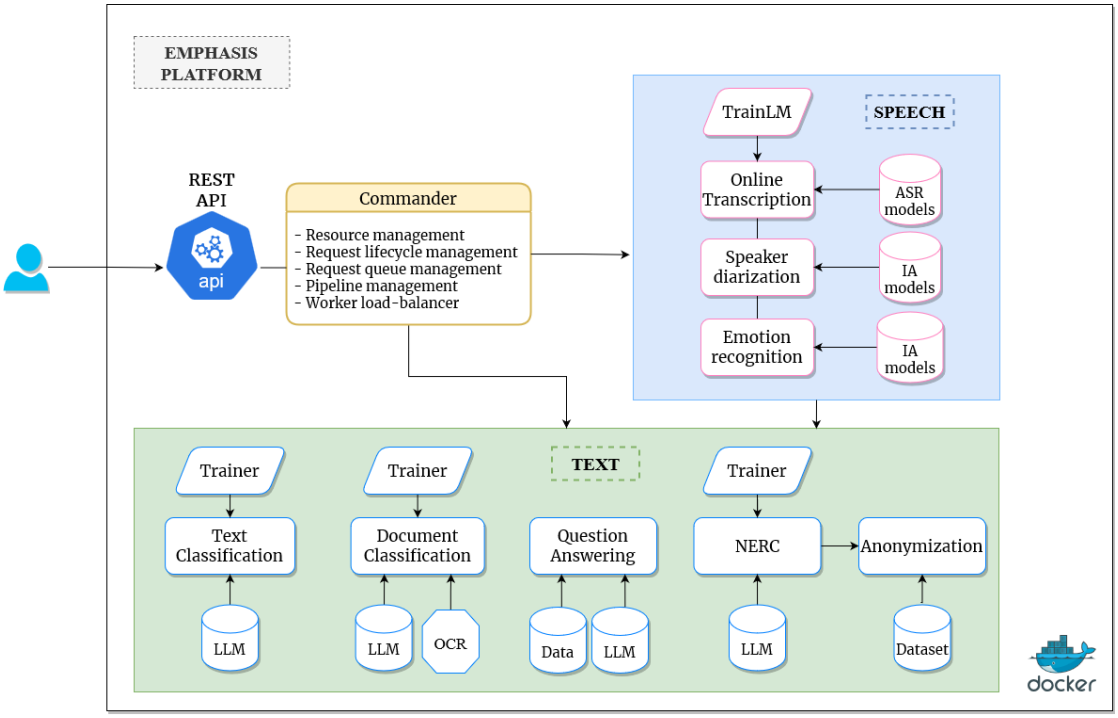

Figure 1: An overview of EMPHASIS platform with the main modules that are integrated with the text and speech technologies developed in the project.

7.1. Text technologies

Text Classification This module focuses on fine-tuning a foundational model (LLM), in this case, BERT [9], to perform text classifications for a set of categories. The training process enables the addition of different sources of texts, and of symbolic features, exploited by the calculation of an embedding layer which is concatenated to the output hidden states of the LLMs. This module can be used for different applications such as sentiment analysis and news classification. The module has two main functionalities: the training of new models and the use of already trained models through a REST API.

Document Classification This module contains the entire document classification pipeline which is developed in two steps. The first one contains the extraction of the text from the documents via document parsing (for text-based documents) or Tesseract OCR (for image-based documents). The second one contains a document trainer used to train a document classifier with different kinds of LLMs such as BERT-based [9] or LayoutLM [10]. The goal of this trainer is to classify documents into a set of labels. It could be done at the page level or the document level. The module has two main functionalities: training new models to classify types of documents, and using trained models to classify documents.

NERC and Anonymization This module contains two main tools. The first one is a sequence-labeller trainer that could be used to train a NERC disambiguator with an annotated dataset at entity level. In this case, we have implemented a trainer that accepts an arbitrary number of entities and training a LLM to recognise and classify entities. Again, this module could be used to train new models or to use it with existing ones. The second tool, named anonimisator, performs the replacement of the detected entities of the NERC tool into a set of candidates based on a taxonomy of entities. This taxonomy is defined in a YML file and performs a set of techniques based on the kind of entity, for instance replacement or obfuscation. Moreover, in the case of entities such as Persons or Locations, a replacement could be done using a predefined dictionary or in the case of a category such as telephone numbers, an algorithm performing a random production of numbers could be used. All these possible settings are easily configurable from the YML file.

Question Answering This module allows the user to get answers to a given question/query automatically. The system is designed to find answers in a database, which consists of a large corpus of texts written in natural language and relevant to each use case. There are three modes of searching: (i) semantic similarity, (ii) extractive QA and (iii) lexical search. The semantic similarity method is implemented with LLMs trained to distinguish semantics [11], so they retrieve paragraphs or phrases most relevant to a question. The extractive QA algorithm uses a language model pre-trained to extract a short span of text with the answer from a given document. Finally, Lexical search includes the algorithms to perform an exact string matching and a statistical ranking based on words. This module has, on the one hand, the functionality of indexing and vectorising a database (from a use-case) in Elasticsearch[10] and, on the other hand, the functionality of fast document retrieval using the modes explained before.

7.2. Speech Technologies

Online Transcription The Online Transcription module aims to convert the input audio into text and it is based on Vicomtech's proprietary Transkit software library. The library offers easy access to speech transcription functionalities through a REST API, supports concurrent processing, can be deployed as a standalone application or in scalable mode with automatic request traffic balancing, and includes dynamic management of decentralised transcription instances.

The library is composed of 7 technological modules connected via configurable pipelines. The modules correspond to an audio transcoder which integrates the FFmpeg[11] tool, an acoustic segmenter based on the Voice Activity Detector (VAD) module proposed by [12], an Automatic Speech Recognition (ASR) module built on top of Kaldi [13], a module for automatic punctuation and capitalisation [14], a rule-based text normaliser, and a final postprocessing module in charge of generating different output formats.

With the aim of providing domain adaptation functionalities, the TrainLM sub-module was developed, which enables the user to adapt the Language Model and Vocabulary of the ASR at text level.

Diarisation Speaker diarisation aims to solve the problem of "who spoke when", which implies segmenting a given audio over the active speaker and clustering the segments belonging to each speaker by assigning the same label to each one [15]. Unlike speaker recognition technologies, no prior knowledge about the speakers' identity or the total number of speakers in the audio is required to perform the diarisation task.

The diarisation module for the EMPHASIS solution was developed on top of the Kaldi toolkit, through a technological approach based on the combination of identity vectors and a posterior classification task, similar to the work presented in [16]. The identity vectors (i.e., x-vectors) are extracted from a time-delay neural network (TDNN) model [17], while the Probabilistic Linear Discriminant Analysis (PLDA) is used to score the similarity between embeddings. In the last step, the VBx resegmentation algorithm [18] is employed to obtain the final sequence of speakers given the speaker-specific distribution scored from the previous PLDA model.

Emotion Detection The Speech Emotion Recognition (SER) module was built with the aim of analysing the emotions of agents and customers in telephone calls from a call centre. To this end, and considering the lack of available data for the Spanish language to train AI models within this complex domain, a new acoustic corpora was collected and annotated for the task within the project from a collection of telephone calls gathered from 2 companies of the consortium, each with its particular domain. This audio data was manually labelled considering the particularities of each domain at categorical and the three-dimensional (arousal, valence and dominance) levels.

Regarding the technological approach, and following the current trends, an architecture composed of a feature extractor (i.e., audio embeddings) module and a classification layer was implemented. As the feature extractor, the Wav2Vec2.0 XLS-R foundation model [19] was integrated, whilst as classification layer both Support Vector Machines (SVM) and DNN based downstream models were evaluated with a very similar performance.

8. Conclusions

The main goal of the EMPHASIS project is to support companies in the hyperautomation of tasks by researching the most convenient tools adaptable to different domains and using the novel state-of-the-art technologies based on deep learning technology. As a result of the project, companies have a platform with plug-and-play core Text and Speech Analysis modules adaptable to different domains and languages.

Acknowledgments

EMPHASIS was partially funded by the Basque Business Development Agency, SPRI, under grant agreement ZE-2021/00039. The authors would also like to thank the companies in the consortium that have contributed with

[10] https://www.elastic.co/es/elasticsearch
[11] https://ffmpeg.org/

their knowledge and experience to the project and the EMPHASIS solution.

References

[1] A. Baidya, Document analysis and classification: A robotic process automation (rpa) and machine learning approach, in: 2021 4th International Conference on Information and Computer Technologies (ICICT), IEEE, 2021, pp. 33–37.

[2] A. Haleem, M. Javaid, R. P. Singh, S. Rab, R. Suman, Hyperautomation for the enhancement of automation in industries, Sensors International 2 (2021) 100124.

[3] A. Álvarez, H. Arzelus, I. G. Torre, A. González-Docasal, Evaluating novel speech transcription architectures on the spanish rtve2020 database, Applied Sciences 12 (2022) 1889.

[4] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown, Text classification algorithms: A survey, Information 10 (2019) 150.

[5] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, Layoutlm: Pre-training of text and layout for document image understanding, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1192–1200.

[6] A. González-Docasal, N. Pérez, A. Alvarez, M. Serras, L. García-Sardiña, H. Arzelus, A. García Pablos, M. Cuadros Oller, P. Delgado, A. Lazpiur, B. Romero, Nalytics: Natural speech and text analytics, 2020-09.

[7] J. M. Martín-Doñas, I. G. Torre, A. Álvarez, J. Arellano, The vicomtech spoofing-aware biometric system for the sasv challenge, arXiv preprint arXiv:2204.01399 (2022).

[8] O. de Gibert Bonet, A. García Pablos, M. Cuadros, M. Melero, Spanish datasets for sensitive entity detection in the legal domain, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3751–3760. URL: https://aclanthology.org/2022.lrec-1.400.

[9] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[10] A. R. GV, Q. You, D. Dickinson, E. Bunch, G. Fung, Document Classification and Information Extraction framework for Insurance Applications, in: 2021 Third International Conference on Transdisciplinary AI (TransAI), 2021, pp. 8–16. doi:10.1109/TransAI51903.2021.00010.

[11] N. Reimers, I. Gurevych, Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020.

[12] H. Dinkel, S. Wang, X. Xu, M. Wu, K. Yu, Voice activity detection in the wild: A data-driven approach using teacher-student training, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 1542–1555.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The kaldi speech recognition toolkit, in: IEEE 2011 workshop on automatic speech recognition and understanding, CONF, IEEE Signal Processing Society, 2011.

[14] A. González-Docasal, A. García-Pablos, H. Arzelus, A. Álvarez, Autopunct: A bert-based automatic punctuation and capitalisation system for spanish and basque, Procesamiento del Lenguaje Natural 67 (2021) 59–68.

[15] T. Park, N. Kanda, D. Dimitriadis, K. Han, S. Watanabe, S. Narayanan, A review of speaker diarization: Recent advances with deep learning, Computer Speech and Language 72 (2022) 101317.

[16] G. Sell, D. Garcia-Romero, Speaker diarization with plda i-vector scoring and unsupervised calibration, in: 2014 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2014, pp. 413–417.

[17] V. Peddinti, D. Povey, S. Khudanpur, A time delay neural network architecture for efficient modeling of long temporal contexts, in: Sixteenth annual conference of the international speech communication association, 2015.

[18] F. Landini, J. Profant, M. Diez, L. Burget, Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks, Computer Speech & Language 71 (2022) 101254.

[19] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.