Quantum Machine Intelligence (2025) 7:46
https://doi.org/10.1007/s42484-025-00260-w

RESEARCH ARTICLE

Multilingual multi-task quantum transfer learning

Giuseppe Buonaiuto[1] · Raffaele Guarasci[1] · Giuseppe De Pietro[2] · Massimo Esposito[1]

Received: 9 September 2024 / Accepted: 12 February 2025

Abstract
Hybrid quantum-classical algorithms have emerged as promising candidates for overcoming current limitations of deep learning techniques and recently have attracted a lot of attention for their application in natural language processing (NLP). Among the potential applications of quantum computing in this field, quantum transfer learning—using quantum circuits for fine-tuning pre-trained classical models specific to a task—is regarded as a potential avenue to exploit the potentiality of quantum computers. This study validates, both experimentally and with domain knowledge analysis, the efficacy of quantum transfer learning for two distinct NLP tasks—semantic and syntactic—and employ multilingual data encompassing both English and Italian. In particular is hereby demonstrated that embedded knowledge coming from pre-trained deep learning models can be effectively transferred into a quantum classifier, which shows good performances, either comparable or potentially better than their classical counterparts, with a further reduction of parameters compared to a purely classical classifier. Furthermore, a qualitative linguistic analysis of the results is presented, that elucidates two points: the lack of language dependence in the quantum models and the ability to discriminate with higher precision than standard classifiers, sub-types of linguistic structures.

Keywords Quantum machine learning · Quantum natural language processing · Variational quantum classifier · Natural language processing · Neural language models

1 Introduction

In this fruitful natural language processing (NLP) season, ruled by neural language models (NLMs), one of the most successful techniques widely used for a large number of tasks is transfer learning (TL). Given the enormous voracity of data for training and fine-tuning the latest neural models of language, based primarily on transformers architecture (Vaswani et al. 2017), the possibility to adapt a model trained for one task to perform a related task has paved the way for endless possibilities.

On the one hand, in the field of NLP, NLMs have become the standard for understanding and generating natural language, allowing for excellent performance in tasks such as text classification, sentiment analysis, and coreference resolution. On the other hand, techniques like TC enable the adaptation of these models by fine-tuning them on more targeted tasks (Buonaiuto et al. 2024; Cardillo et al. 2024; Zaman-Khan et al. 2024; Guarasci et al. 2023). This process exploits models' general language knowledge and refines it in order to achieve better performance on specific tasks, even with limited task-specific data.

Many tasks have benefited from TL (Ruder et al. 2019), ranging from cross-lingual approaches to address low-resource language limitations (Schuster et al. 2018; Guarasci et al. 2021) and inferring linguistic knowledge across typologically different languages (Guarasci et al. 2022; Kim et al. 2017). It also has proven to be helpful in domain adaptation (Ma et al. 2019), machine translation (Shah et al. 2018), and named-entity recognition tasks (Ruder et al. 2017).

In the wake of the recent research that aims to overcome the limitations of current approaches by exploiting quantum computing properties, interest in quantum machine learning (QML) has emerged. Among the possible fields of application of QML, the newly established research field of quantum NLP (QNLP) (Coecke et al. 2010) is fast growing for its

✉ Raffaele Guarasci
raffaele.guarasci@icar.cnr.it

[1] National Research Council of Italy (CNR), Institute for High Performance Computing and Networking (ICAR), Naples 80131, Italy

[2] Department of Information Science and Technology, Pegaso Telematic University, Naples, Italy

Published online: 26 March 2025

potential applications in linguistics. QNLP aims to boost performance in language-related tasks using quantum properties or quantum theory-based algorithms and experimenting with real quantum hardware.

Nevertheless, quantum algorithms must face several problems when deployed (Schuld et al. 2021) mainly related to scalability, hardware limitation, and noise reduction (Ren et al. 2022; Li and Deng 2024). These factors hinder the quality of the models and limit their applicability to real case scenarios. A possible solution is represented by hybrid approaches combining classical pre-trained models with quantum techniques (Li et al. 2022, 2023). Such methods integrate specific model layers on quantum devices, while classical models handle either non-linear operations or the whole optimization process. This is de facto a quantum version of the classical transfer learning algorithm, where knowledge extracted from a model trained on large or general-purpose dataset is then used to train a smaller model (possibly together with a subset of the original model) for handling a specific problem.

Starting from the approach proposed in Li et al. (2023), this work explores the potential of quantum-inspired transfer learning to address the challenge of achieving relevant performances, specifically classification accuracy across diverse tasks and languages in natural language processing (NLP). This work's primary objective is to investigate this hybrid framework's efficacy in capturing and transferring knowledge relevant to NLP tasks. The aim here, in particular, is to elucidate whether such an approach shows some dependency on language-specific factors in unraveling its effectiveness. Furthermore, an attempt has been made to understand if quantum classifiers can adapt to different NLP subtasks. To achieve a comprehensive understanding, the experiments have been conducted employing multilingual datasets, both for English and Italian. Datasets are further expressions of the two distinct NLP subtasks considered: **sentence acceptability judgments** (focusing on syntax) and **sentiment analysis** (focusing on semantics). This multi-task exploration utilizes established corpora, ItaCoLa (Trotta et al. 2021) and SentiPolc (Basile et al. 2014) for Italian, and CoLA (Warstadt and Bowman 2019) and SST-2 (Socher et al. 2013) for English. Notice here that the presented results are both an extension and a further deep and complete investigation of the seminal analysis, constructed by the same author, of quantum transfer learning limited to acceptability judgment on the ItaCola corpus (Buonaiuto et al. 2024). By analyzing the performances of each model selected across the diverse datasets and tasks, together with a linguistical and comparative analysis of the results, encompassing various types of sentence structures, the generalizability of the approach for the NLP case studies and languages is assessed.

1.1 Aim and contribution

The main goal of this paper is to systematically evaluate the impact of quantum transfer learning in a multilingual, multi-task scenario in the QNLP field. For such reasons, two different classification tasks are considered, one syntactic and one semantic, in Italian and English, respectively, using datasets widely used in NLP to evaluate these tasks. In this respect, the following contributions are provided here:

- Assessing learning capabilities of variational quantum circuit (VQC) in both semantic and syntactic NLP tasks in different languages
- Setting a baseline to measure current algorithms' performance and possible future directions, both in terms of the number of parameters and absolute performance compared to classical methods
- Systematically evaluate the results quantitatively using well-known metrics from the literature and compare them with state-of-the-art NLMs
- Carrying out a qualitative multi-task, multi-language comparison to understand what impacts linguistic features and how in terms of explainability

1.2 Outline

The rest of the paper is organized as follows. Section 2 reviews the most recent related work on QML and QNLP in general. Section 3 shows the experimental assessment, including the datasets, tasks, and models under consideration. In Section 4, methodological details for constructing an algorithmic comparison between a quantum and classical pipeline for classification are outlined, while results are presented in Section 5, both from quantitative and qualitative perspectives. Finally, concluding remarks and possible future directions are hinted at in Section 6.

2 Related work

In recent years, quantum mechanics has enhanced preexisting machine learning (ML) algorithms. This fellowship has produced results such as quantum support vector machines (QSVMs), a quantum-inspired genetic algorithm (QGA), and quantum-inspired particle swarm optimization (QPSO). Early attempts of quantum language models (QMLs) have also been proposed, exploiting quantum probability theory to model natural language (Basile and Tamburini 2017; Chen et al. 2021).

Another family of quantum-inspired approaches, quantum neural networks (QNNs) and quantum Boltzmann machines

(QBMs), is particularly effective in tasks such as text classification.

The most successful approaches to date have been those based on quantum transfer learning (Mari et al. 2020), namely, the quantum realization of transfer learning algorithms. Transfer learning is a well-established technique in machine learning (ML) that reuses a model initially developed for a specific task, which is then adapted to a related task. The idea behind transfer learning is using prior knowledge derived from a previous task as a starting point to improve learning and performance on a new task, usually with a small amount of data. Given its versatility, transfer learning has been used on many heterogeneous NLP tasks (Ruder et al. 2019). In particular, it has proven to be suitable to improve performance in named-entity recognition (Ruder et al. 2017), domain adaptation (Ma et al. 2019), machine-translation (Shah et al. 2018), inference (Guarasci et al. 2022; Kim et al. 2017), and cross-lingual approaches (Schuster et al. 2018; Guarasci et al. 2021).

Mutatis mutandis QTL can be implemented in two ways: on the one hand, quantum-transfer learning algorithms that use feature vectors extracted from a trained quantum machine learning algorithm and then fed into a quantum neural network, and on the other, classical-quantum transfer learning algorithm (Mari et al. 2020), an approach encoding input features extracted from a classical network in a multi-qubit state, then computed using a quantum circuit. Output probabilities are projected to the task label space, and parameters are updated using losses.

The natural application implication of QTL techniques lies in QNLP tasks. QNLP leverages quantum mechanics for language data analysis. Leaving aside early theoretical approaches, proposing quantum algorithms without real data testing (Zeng and Coecke 2016; Coecke et al. 2020), approaches in line with the purpose of this work have tested on real datasets using either classical hardware (quantum-inspired) or current quantum machines (quantum-computer). Quantum-inspired approaches integrate quantum mechanics advancements into classical models to enhance performance. Quantum-computer approaches are limited by current hardware to small-to-medium datasets, particularly in classification tasks (Guarasci et al. 2022).

Hybrid approaches combine classical and quantum techniques to overcome scalability issues (Grant et al. 2018; Callison and Chancellor 2022). A notable example is the quantum self-attention neural network (QSANN) (Li et al. 2022), though it faces challenges with hardware switching. A promising solution is the classical-quantum transfer learning paradigm (Mari et al. 2020), which uses pre-trained quantum encodings for scalable QNLP models, paving the way for implementation on real quantum hardware (Li et al. 2023). For a detailed review of QNLP approaches, refer to Guarasci et al. (2022).

3 Materials and methods

3.1 Quantum transfer learning and variational quantum circuits

While in standard AI, transfer learning (Weiss et al. 2016) is commonly used for exploiting the capabilities of large pre-trained deep networks, it only recently matters the construction of its quantum counterpart (Mari et al. 2020). While various methods are in place for realizing quantum transfer learning protocols, the most widespread approach in use—that is, the one implemented in this work—consists of an application of quantum neural networks, i.e., parametrized quantum circuits, on the extracted featured vectors of pre-trained model for realizing the specific tasks, either classification or regression. In NLP, it is customary to use pre-trained transformer-based NLMs for representing language, either text, speech, or both, in an expanded algebraic vector space that entails both the single word representations and their positional value, thus their role within the semantics of the phrase.

Here, the quantum transfer learning models are constructed following precisely the scheme mentioned above: NLMs such as Bert and Electra are used to extract feature vectors (embeddings) from the different corpora used in both languages considered (English and Italian). These embeddings are then encoded via a suitable feature map into quantum states and fed into a variational quantum circuit (VQC) trained to perform the specific task. In this way, leveraging the knowledge from the pre-trained classical model, the quantum model learns the patterns within the date, possibly fostering a faster convergence and greater expressivity of classical deep learning models, given the richness of the Hilbert space vectors exploited during the computation.

Formally, a VQC of a given depth d is a series of stacked quantum layers, i.e., of unitary operators U_i, with $i = 0...d$, composed of quantum gates, some of which are parametrized θ_i^j, being j the index of the parametrized gate belonging to layer i. These series of quantum layers act on the initial quantum state $|s\rangle$, representing the encoded embeddings generated via the pre-trained LLM in use. In a compact way:

$$|y\rangle = \left(\bigotimes_{i=1}^{d} U_i(\theta_{\mathbf{i}}) \right) |s\rangle \tag{1}$$

where $|y\rangle$ is the final state and $\theta_{\mathbf{i}} = (\theta_i^0 ... \theta_i^n)$ is the set of the n parameters in the layer i. The final state is then measured via

a set of measurement operators to obtain a real number that is used to quantify the objective function $\mathcal{L}$ on the specific problem to be solved, i.e.,

$$\mathcal{L}(y_{true}, \hat{y}) \quad with \quad \hat{y} = \langle M \rangle_{|y\rangle} = \frac{1}{N} \sum_{i=1}^{N} \langle y_i | M | y_i \rangle \tag{2}$$

where the variational circuit is evaluated N times to estimate the expectation value of the measurement outcome. The objective function value is then used, as in the classical supervised learning protocol, to update, via gradient descent, the parameters of the quantum circuit until convergence is reached.

3.1.1 Classical to quantum data encoding

As stated above, the first step necessary to realize a quantum transfer learning pipeline consists of the initial state preparation. It is a computationally expensive step but rather fundamental to exploit the in-full VQC: the quantum computation is, in fact, based on quantum state vectors. Hence, classical data, which are not quantum by nature, need to be appropriately encoded into quantum states, possibly minimizing the information lost in the process. While several approaches exist to implement this step, it is still an open and debated research field: encoding classical data into a quantum state is non-trivial (Barnum et al. 2001) and requires quantum resources to be effective. In this regard, a series of methods exist, ranging from quantum kernels, Blank et al. (2020) to circuit-based encoding (Park et al. 2019), which attempt to make the translation effective.

In this work, as aforementioned, the strategy in use is the **amplitude embedding**. With this strategy, the classical embedding vectors coming from the processing of the dataset through the pre-trained NLMs are encoded into the amplitudes of a superposition of n qubits. As each qubit is spanned in a two-dimensional complex Hilbert space, a superposition of n qubit is composed of at most 2^n components; hence, it is possible, with the amplitude embedding protocol, to embed large classical vectors using relatively few qubits. Specifically, suppose the classical embedding $\mathbf{x}$ has dimension D. In that case, the qubit required to realize the state preparation with amplitude embedding is $n = \lfloor \log_2 D \rfloor$, i.e., the greatest integer close to the dimension's $\log_2$. Suppose an excessive number of qubits is necessary concerning the required dimensions of the embedders. In that case, the remainder amplitude is padded to fully realize the 2^n feature vectors. It is necessary to normalize the classical embeddings before using them to realize the quantum state, as the probabilistic interpretation of quantum amplitudes needs to hold, i.e.,

$$\sum_{k=1}^{2^n} \frac{x_k}{||\mathbf{x}||} = \sum_{k=1}^{2^n} \alpha_k = 1, \tag{3}$$

so that the modulus squared of a component, say $||\alpha_i||^2$ represents the probability of measuring that value based on the quantum state. Once the embedding is normalized, a set of CNOT and rotation gates is applied (following the Motten state preparation scheme Möttönen et al. 2005) to encode the data into a quantum state.

3.2 Tasks and datasets

3.2.1 Acceptability judgments

Acceptability judgments (henceforth AJ) are a crucial task that has interested computational and theoretical linguistics scholars since the early days. Although assessing how correct a sentence sounds may seem trivial to a native speaker, there are numerous open issues that the proposed approaches have faced. Among the various criticisms, the most obvious is the subjectivity and context-dependence of such judgments, which are deeply affected by fine-grained linguistic features ranging from syntax to semantics. Therefore, the reliability of AJ as a source of linguistic data has often been questioned (Linzen and Oseki 2018).

With the rapid growth of increasingly refined language models and the consequent creation of more extensive resources, there has been a marked leap forward with respect to this task. In particular, architectures based on deep neural networks (RNNs, CNNs, or transformers) have proven to be able to exploit syntactic structure and semantic content to accurately AJ prediction.

A large part of the credit for the renewed interest in the task can be attributed to the released corpora, starting with that for the English language (Warstadt et al. 2019). Similar criteria have been adopted to create other resources covering different languages. So far, resources have been developed for Russian (Mikhailov et al. 2022), Japanese (Someya et al. 2023), Norwegian (Jentoft and Samuel 2023), Swedish (Volodina et al. 2021), Spanish (Bel et al. 2024), and Italian (Trotta et al. 2021).

More recently, a multi-language approach to perform a comparative evaluation of different NLMs on such a dataset has been proposed (Zhang et al. 2023). In this work, ten languages belonging to different families (germanic, romance, slavic, sino-tibetan, japonic, and Semitic languages) have

been taken into account using existing resources or proposing resources built from scratch *ex novo*.

For the purpose of this work, the datasets used are the following:

- The **Corpus of Lingusitic Acceptability** (CoLa) (Warstadt et al. 2019), for the English language: It is the first large-scale dataset for acceptability judgments. Sentences are binary labeled and extracted from various linguistic literature. It has become such a popular resource as to be included in GLUE benchmark (Wang et al. 2018), a very popular multi-task benchmark for English natural language understanding. It is composed of 9594 sentences.
- The **Italian Corpus of Linguistic Acceptability** (ItaCoLa) (Trotta et al. 2021; Bonetti et al. 2022): This corpus is composed by approximately 9700 sentences drawn from various sources that cover numerous linguistic phenomena. Expert linguists label every sentence as acceptable (1) or unacceptable (0). Sentences are extracted from various sources in order to represent a wide spectrum of linguistic phenomena (Table 1).

3.2.2 Sentiment analysis

Sentiment analysis, which can be further subdivided into polarity and emotion detection, is an NLP classification task that aims to extract the sentiment expressed in a natural language text (Birjali et al. 2021). Given this work's purpose, only the sentiment analysis version formalized as a binary classification task is considered. Thus, the goal is to label a text as positive or negative based on some lexical elements contained in it or specific syntactic structures (i.e., negation) that convey a precise semantic value (Dai et al. 2022).

Sentiment analysis approaches can be divided into three main types (Al-Qablan et al. 2023). Early approaches focused on lexicons and rules, exploiting handwritten resources annotated by experts (Khoo and Johnkhan 2018). A second strand includes automatic approaches based on machine and deep learning (Naresh et al. 2021; Rani and Kumar 2019). Finally, a family of approaches very successful in recent years is hybrid ones (Stacked Machine Learning 2020), trying to combine the scalability of automatic techniques with rule-based expressive power and accuracy in specific use cases.

Concerning the work presented here, sentiment analysis datasets taken into account are the following:

- **Stanford Sentiment Treebank** is a corpus with complete labeled dependency parse trees (DPTs) created for the specific purpose of analyzing sentiment. The corpus collects 11,855 from movie reviews, and it is based on the dataset introduced by Pang et al. (2008). Notice that, for this work, DPTs have not been used but sentences in plain text, having no counterpart in the other language.
- The **SENTIPOLC (sentiment polarity classification)** dataset is an annotated collection used for sentiment analysis and opinion mining in Italian. It has been introduced within the evaluation campaign for the shared task of sentiment analysis in Italian tweets (EVALITA) (Basile et al. 2014). Texts are extracted from Twitter.

3.3 Models

For this work, two neural language models (NLMs) have been considered, namely Bert and Electra.

3.3.1 Bert

Among NLMs in literature, Bert (Devlin et al. 2019) is the most widely used due to its efficiency and high performance. Generally speaking, Bert is a multi-layer bidirectional architecture based on the original transformer encoder (Vaswani et al. 2017), pre-trained on large-scale unlabeled text via two training goals, i.e., masked language modeling and next sentence prediction.

A pre-trained Bert model typically provides a powerful context-dependent sentence representation that can be successively adapted to a downstream NLP task through a fine-tuning procedure according to different needs. The fine-tuning procedure requires configuring several hyperparameters whose values directly influence the results that can be obtained.

The $Bert_{base}$ model consists of 12 hidden layers, each one having 768 hidden dimensional states and 12 attention heads, with a total of 110 M parameters. The $Bert_{base}$ model accepts input sequences of words with a maximum length of 512. Each model layer encodes a distinct embedded representation

Table 1 Overview of the datasets used for two different tasks

Dataset	Task	Language	Size	Source	Annotation
CoLa	Acceptability judgements	English	9594	Linguistic literature	Domain experts
ItaCoLa		Italian	9700		
SentiPolc	Sentiment analysis	English	6421	Twitter	
STN-2		Italian	11,855	Movie reviews	

Size is expressed in number of sentences, eventually including all the different splits (train, test, and dev)

of the input words, which can be leveraged for various NLP tasks, including the syntactic probe discussed in this paper.

Masked language modeling involves randomly masking a percentage of words in the training corpus. By doing so, the pre-trained model learns to encode information from both directions of the sentences and simultaneously predict the masked words. The input vocabulary can be either *cased* or *uncased*, resulting in two different pre-trained models. The flexibility offered by bidirectional analysis simultaneously allows, on the one hand, to maintain a large generating capacity through the inner layers of the deep constituent network and, on the other hand, to use the outer layers of the network to adapt to the specific task through the fine-tuning phase, is what has allowed Bert to be the benchmark model in the literature in recent years.

Bert expects that each input sequence of words starts with a unique token *[CLS]*, used to obtain in output a vector of size H, i.e., the size of the hidden layers, representing the entire input sequence. Moreover, the unique token *[SEP]* must be placed within the input sequence at the end of each sentence.

Given an input sequence of words $\eta = (\eta_1, \eta_2, ..., \eta_m)$, the output of Bert is $h = (h_0, h_1, h_2, ..., h_m)$ where $h_0 \in R^H$ is the final hidden state of the special token *[CLS]* and provides a pooled representation for the full input sequence, while h_i are the final hidden states of other input tokens.

To fine-tune Bert for classifying input sequences of words into Y different text categories, the final hidden state h_0 can be used to feed a classification layer, with a subsequent softmax operation to turn the scores of each text category into likelihoods (Sun et al. 2019):

$$P = softmax(CW^T) \tag{4}$$

where $W \in R^{YxH}$ is the parameter matrix of the classification layer.

3.3.2 Electra

The second NLM taken into account is Electra (Clark et al. 2020), since it has shown a better ability to capture contextual word representations outperforming, in its downstream performance, other models, like Bert, given the same model size, data, and compute (Rogers et al. 2020).

Generally speaking, Electra is a pre-training approach that trains two transformer models, namely the generator G and the discriminator D. The role of the model G is to replace tokens in a sequence and is, therefore, usually trained as a masked language model. The model D, which is typically the Electra model of interest, tries instead to identify which tokens were replaced by G in the sequence, and it may be a Bert-based model, virtually any model producing an output distribution over tokens.

In particular, for a given input sequence, where some tokens are randomly replaced with a special *[MASK]* token, G is trained to predict the original tokens for all masked ones, after which G generates a fake input sequence for D by replacing the [MASK] tokens with fakes. Finally, D is given the fake sequence as input and is trained to predict whether their tokens are original or *fake*. This approach, replaced by token detection (RTD), allows the use of a minor number of examples without losing performance.

More formally, given an input sequence of token extracted from a raw text $s = w_1, w_2, \ldots, w_n$ being w_t $(1 \leq t \leq n)$ the generic token, both G and D firstly encode s into a sequence of contextualized vector representations $h(s) = h_1, h_2, \ldots, h_n$.

Then, for each position t for which $w_t = [MASK]$, the generator G predicts, through a softmax layer, the probability to generate a specific token w_t:

$$p_G(w_t|s) = \frac{e(w_t)^T h_G(s)_t}{\sum_{w'} \exp(e(w')^T h_G(s)_t)} \tag{5}$$

where $e(\cdot) : w_t \in s \rightarrow R^{dim}$ is the embedding function, and dim the chosen embedding size.

The discriminator D predicts, via a sigmoid layer, if w_t is original or "fake":

$$D(s,t) = sigmoid(e(w_t)^T h_D(s)_t) \tag{6}$$

During the pre-training, the following combined loss function is minimized:

$$\min_{\theta_G, \theta_D} \sum_{s \in \chi} \mathcal{L}_{Gen}(s, \theta_G) + \lambda \mathcal{L}_{Dis}(s, \theta_D) \tag{7}$$

where $\mathcal{L}_{Gen}$ and $\mathcal{L}_{Dis}$ are the loss functions of G and D, respectively.

At the end of the pre-training, G is discarded and only D is effectively used for fine-tuning on the specific task.

Masked language modeling pre-training methods such as Bert corrupt the input by replacing some tokens with [MASK] and then training a model to reconstruct the original tokens. While they produce good results when transferred to downstream NLP tasks, they generally require large amounts of computing to be effective. As an alternative, replaced token detection is a more sample-efficient pre-training task that corrupts the input by replacing some tokens with plausible alternatives sampled from a small generator network instead of masking the input. The main reason Electra efficiency results improved concerning Bert-like NLMs is that predictions are calculated not only over masked tokens but also for the other tokens in the input sequence, and, thus, the discriminator loss can be calculated over all input tokens. It allows

using a minor number of examples without losing in performance.

4 Computational pipelines

In this section, the methodological details for constructing a quantum pipeline that learns to classify a text given an embedded representation of the various datasets are outlined.

Using the approaches outlined in previous sections, entries of the vector embeddings generated by Bert or Electra, on both languages, are encoded in the amplitude of a superposition state of the $N = \log_2 n$ qubits in use, via amplitude embedding, such that, measuring the probabilities (the modulus square) of each component of the quantum state vector hitherto constructed gives back the original embedding representation. The variational quantum ansatz is then used to process the quantum state vector generated. The function of the variational ansatz is somehow equivalent to classical multilayer perceptrons: parameters are iteratively updated, subject to the evaluation of an objective function, which needs to be minimized, associated with the problem to be tackled. The output of the variational quantum circuit is then measured to extract the information out of the quantum states. The measurement hereby used is represented by $Pauli - Z$ operators, which corresponds to the projection of the qubit state (which can be represented as a Bloch sphere in the Hilbert space), on the Z axis. At the completion of the training phase, the performances of the binary classifier are measured via the accuracy. The quantum pipeline for the binary classification in the contexts of acceptability and sentiment analysis, for both English and Italian (Guarasci et al. 2024), is compared with the classical counterpart, i.e., the embeddings are processed by a multi-layer perceptron that performs the classification: this step allows to compare the effectiveness of the quantum classifier and to address, via a qualitative analysis, what the quantum circuit is actually learning, which means how it weights the different elements of the sentences in the final scoring for labelling the class.

4.1 Variational ansatz and entanglement

As stated in the previous section, after encoding the embedding vectors in the quantum state amplitudes, the quantum states obtained, representative of the dataset, are processed by a variational quantum circuit. While the parameters are being updated iteratively via a quantum version of the gradient descent (Rebentrost et al. 2019), provided the evaluation of an objective function, the structure of the quantum circuit itself, i.e., the ansatz for the unitary operator that limits the search space into a sector of the Hilbert space, has to be defined.

Several proposals for the optimal ansatz can be found in the literature. In other cases, the ansatz is general, as it applies to various scenarios. It is a general rule (Díez-Valle et al. 2021) to insert a certain amount of entanglement, i.e., nonclassical correlations, among qubits involved in the computation, to exploit the learning capabilities of quantum circuits better: while there is no general rule about the best form of connection between qubit to maximize the performances, some good practices involve taking into account the problem complexity and the topological properties of the quantum hardware in use (Tilly et al. 2022; Buonaiuto et al. 2024).

Here, the variational circuit ansatz in use is the **Strong Entangling Ansatz**, whose function is provided by Pennylane (Bergholm et al. 2018). The ansatz comprises single qubit three-dimensional rotational gates and a ring of CNOT gate connecting a qubit with the next within a specified range. Repeated layers with the same structure can be concatenated and applied to the initial state. In particular, in the following, the ansatz is made of 6 layers with alternating one and two ranges of connection, meaning that the CNOT are connecting the closest qubits in the odd layers and the next nearest qubits on the even layers, as shown in Fig. 1.

5 Results and discussion

In this section, the results of the quantum transfer learning pipeline for the acceptability judgment and sentiment analysis, both on English and Italian corpora, are presented and discussed. In the first part of the section, a quantitative analysis is carried out, describing and discussing the results of the learning strategies, focusing on the performances and the structural details of the learning process.

Both the classical and the quantum classifiers are trained with a batch size of 32. Further, training parameters are the optimizer, the *AdamW* for every algorithm, and the learning rate 10^{-5}, which is fixed for every model in use. The batch size and the learning rate have been chosen after careful hyperparameter optimization. Concerning the embedding extractors, i.e., the pre-trained NLMs Bert and Electra, for each model, the maximum word length has been determined a priori by investigating the entries of each dataset in use.

The classical pipeline is basically composed of a multilayer perceptron, which has the embedding vectors of length 768 as inputs, then passes them through a hidden layer of size 256, evaluates a *ReLu* function, and then again another linear layer of neuron of dimension (256, 2), where the last number indicates the number of classes.

The quantum pipeline, as aforementioned, is composed of an amplitude embedding module for converting the real vectors into quantum states, with 10 qubits for encoding the entire embedding vector, two of which are padded to zero.

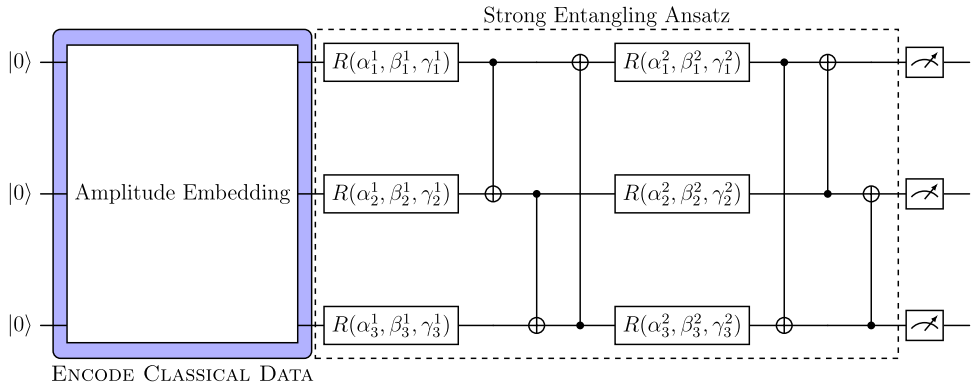

Fig. 1 A schematic representation of the quantum pipeline used for transfer learning. Sentences from the various datasets are first tokenized and then pre-trained models, Bert and Electra, trained either on English or Italian, are used to extract embeddings. These real vector representations of words and sentences are encoded in a quantum state via amplitude embedding and then processed via parametrized quantum circuits, with a strong entangling ansatz structure. In the scheme, a single layer of range 2 of the ansatz is highlighted. The results of the measurement on the quantum states are then used to evaluate the performances of the learning models (i.e., for the binary classification tasks)

Further, 6 layers of strong entangling ansatz are used for training. The output is obtained via a Z measurement on the first qubit, which estimates the data point belonging to one of the two classes. Specifically, if the final value $\langle\psi|Z|\psi\rangle \geq 0.5$, then the point is labeled as class 1, otherwise 0, where $|\psi\rangle$ is the final state generated upon the application of the variational quantum circuit on the initial encoded state.

Notice here that all the models have been constructed in Pytorch. The quantum circuits have been constructed using the built-in Pennylane function, while the exact quantum simulations were carried out using the specification from the backend *IBM Kyev*; hence, its connectivity properties and basis gate have been taken into account. In particular, for each instance of the quantum circuit, a 2000 run of measurements has been performed to collect enough statistics to estimate the mean value of the Z measurement operator. It is worth pointing out that, as prescribed in Buonaiuto et al. (2024), the quantum experiments have been realized minimizing the circuit depth required and selecting sectors of the quantum hardware with smaller gate noise: in this way, the barren plateaus, i.e., the vanishing gradient in the training of a parametrized quantum circuits, were avoided.

5.1 Quantitative analysis

The classical pre-trained models used in the experimental phase, whose results are described in details below, are the following, all readily available online (Wolf et al. 2019):

- **Bert for English language:** *google-bert/bert-base-cased*
- **Bert for Italian language:** *dbmdz/bert-base-italian-xxl-cased*
- **Electra for English language:** *google/electra-base-discriminator*
- **Electra for Italian language:** *dbmdz/electra-base-italian-xxl-cased-discriminator*

Each dataset has been divided into training set, validation set, and test set, as customary, with a ratio of 60%-15%-25% respectively. The split portions of the dataset are randomly selected for each experimental realization. Every experiment has been realized 12 times, in order to estimate the means and the standard deviation of the metrics of interest (Fig. 2).

Acceptability judgement The first batch of experiments concerns the acceptability judgements on both languages: the mean and standard deviation of the accuracy on the validation set for each epoch is shown in Fig. 3 for the Cola dataset. It can be noticed that embeddings generated via Bert, similarly to what happened using Electra, give higher scores in accuracy in the classical case than with the proposed quantum scheme. This fact is strictly related to the combination of the expressivity of the network in use and the quality of the embeddings. Even if relatively simple, the multilayer perceptron possesses more parameters than the variational quantum circuit. While this fact can, in principle, hinder the training, for instance, producing over-fitting, it is, in fact, more powerful for spanning a larger parameter space compared to a model with fewer parameters. This argument by itself is not enough, but it coherently adds up when taking into account the representation power of Bert compared to Electra. The latter, as confirmed by the vast literature on classical transfer learning for NLP, can capture a richer set of hidden features within the dataset, giving rise generally to higher performances. In this sense, when using Electra, the

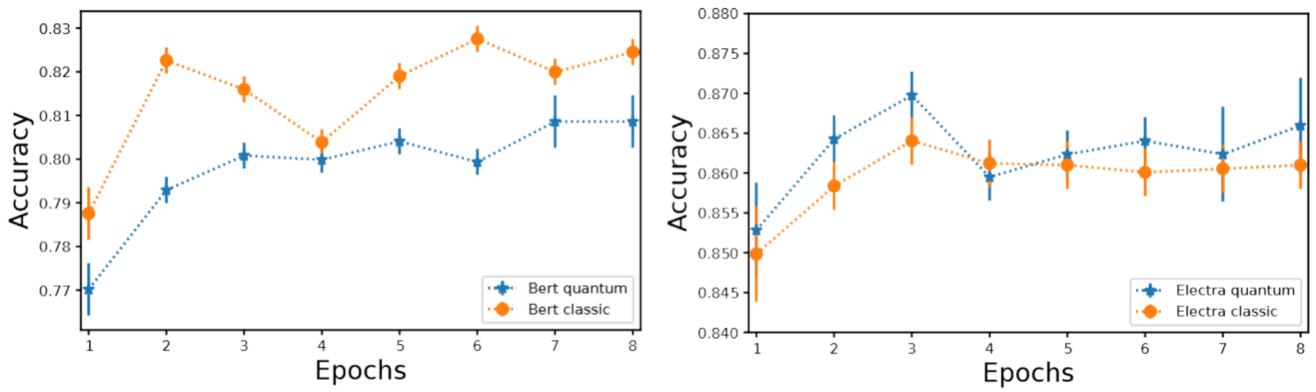

Fig. 2 **Accuracy on validation set** per epoch for Bert and Electra, with either classical or quantum classifiers, for the acceptability judgement performed on the Cola corpus

quantum circuits are facilitated when estimating the correlations (modeled by the entanglement structure of the ansatz) between each term of the embeddings by the robust baseline provided by the pre-trained model.

This fact is confirmed by the behavior of the accuracy on the validation set: mean values for each epoch are generally slightly higher for Electra quantum, compared to the classical one, and the convergence is rapidly reached. It is worth pointing out here that given the nature of the specific task, i.e., acceptability judgment, which is strongly related to the syntactical structure of the sentence, there is a potential space for improvement for the quantum algorithm, provided a meaningful representation of the underlying grammatical structures. It has been demonstrated in fact (Guarasci et al. 2022) that transformer models, such as Bert and Electra, during training, are able to learn some sort of syntactic relationships within each word of a sentence: while these are not necessarily the most correct form of grammars, they represent a learning scheme for constructing representative embedding. A correct form of grammar can be, however, encoded in the quantum circuit via a tensor representation, as demonstrated by Lambek (2006): this kind of information encoding is expected to foster more accurate results for syntactic-related tasks while reducing the size of the training dataset necessary to train the models efficiently. Notice here that the performance obtained for Cola, although scaled given the differences within the dataset, aligns with those given by ItaCola.

Sentiment analysis The second batch of experiments is performed on the sentiment analysis task in both languages. The difference between the present problem and the acceptability lies in the importance of the grammatical structure for the latter, which is almost fundamental in the definition of an acceptable sentence. At the same time, it is secondary, although necessary for the former. The experiments are carried out the same way as before, where the sentiment classification of a sentence, labeled as 0 if negative and 1 if positive, is essentially binary. Results for the mean and standard deviation of the accuracy per epoch on the validation set for SST-2 are shown in Fig. 3: in this case, the accuracy shows a tendency to be higher when a quantum classifier is used, both for Bert and Electra. It is worth clarifying here that the tendency is statistically robust, as shown by the error

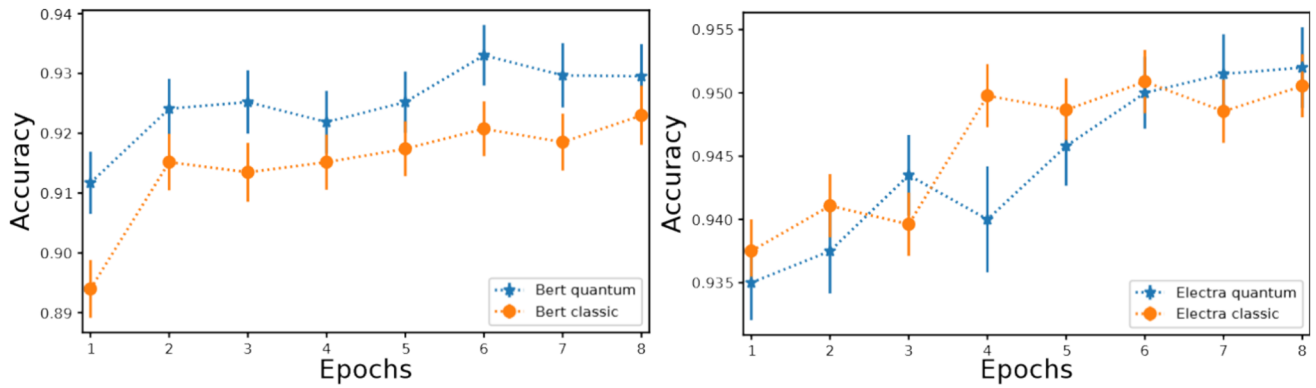

Fig. 3 **Accuracy on validation set** per epoch for Bert and Electra, with either classical or quantum classifiers, for the sentiment analysis performed on the SST-2 corpus

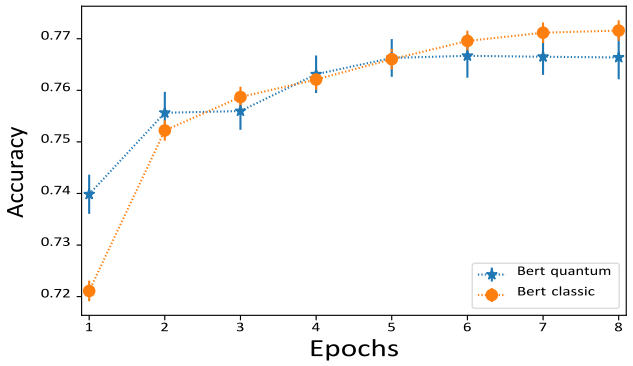

Fig. 4 Accuracy on validation set per epoch for Bert, with either classical or quantum classifiers, for the sentiment analysis performed on the Sentipolc corpus

evaluated for each epoch, even though it is not necessarily a significant deviation from the classical results. It is hence not possible to conclude that, in general, the quantum algorithms outperform the classical ones. At the same time, it is valid that, given the present set of parameters and model specifications, the advantage of the performances of the quantum pipelines is evident. Notice that the quantum classifiers give better results on both languages, as shown in Fig. 4 for Bert embeddings on Sentipolc. The order of magnitude of the performances for the Italian case, compared to the English one, needs to be attributed to the dataset's features. Sentipolc, in fact, is smaller in size compared to SST-2, and the tweets are not pre-processed, meaning that hashtags, emoticons, and URLs are in place, which hinders the quality of the classification. Nevertheless, the models' behavior is consistent with the results obtained on SST-2, with Bert quantum and Bert classical showing similar behavior in validation and Electra quantum performing slightly better than the remainders.

Test set performances and models size Results of the accuracy on the test set, with the relative standard deviation, can be found in Table 2: as expected from the analysis on the validation per epochs, while on the acceptability judgment both on English and Italian, Bert quantum falls shortly behind Bert classical, while Electra quantum and classical give analogous results, the results on the sentiment analysis confirm that Electra quantum gives the best results. This fact goes together with Bert quantum outperforming Bert classical on every test set for the sentiment analysis task. The result suggests that, in line with those mentioned above for the acceptability, the quantum classifier is more adequate than the classical one for classifying vectors on tasks poorly related to the syntactic structure: the quantum circuits involved, in fact, even if not strikingly, always gives better results for sentiment analysis, given the same model structures and hyper-parameters.

Table 3 Number of parameters and average computation time (in seconds) on a single GPU for both types of models in use, i.e., quantum and classical

Model type	#Parameters	Avg. Comp. time
Classical MLP	$\approx 2 \times 10^5$	$\approx 2 \times 10^2 s$
Quantum Circuit	$\approx 3 \times 10^2$	$\approx 4 \times 10^2 s$

These results allow to construct two primary considerations: that encoding the grammar into the quantum state via the quantum natural language approach might constitute an advantage for syntactic-related tasks and that soon, with readily available quantum devices, quantum transfer learning might help in improving the performances of the models while reducing the number of parameters, i.e., the computational resources required for an efficient training.

The considerations mentioned above are shown synthetically in Table 3: the quantum model has, on average, 10^3 fewer parameters than those used in the multilayer perceptron. However, the average computational time for a single epoch on a classical GPU doubles in the quantum case. It is an obvious consequence of the qubit vector representation: the quantum circuit to be evaluated, in fact, is a matrix of dimension $2^n \times 2^n$, where n is the number of the qubits. Hence, its deployment on classical hardware becomes more expensive the greater the number of qubits involved. The computation would have possibly been faster if real quantum hardware had been used, apart from the quantum errors that need to be mitigated.

5.2 Qualitative analysis

As shown in the Table 2, it is possible to see a disparity in behavior between the semantic and the syntactic tasks. Regarding NLMs, Electra proves to be more efficient than Bert in all cases, regardless of the task and language under consideration. This result aligns with several other studies

Table 2 Comparison of classification accuracy on different datasets using a classic approach based on Bert and Electra and the corresponding quantum transfer learning pipelines

Dataset	Bert classical	Bert quantum	Electra classical	Electra quantum
Cola	0.815 ±0.008	0.795 ±0.008	**0.842** ±0.005	**0.842** ±0.005
ItaCola	0.904 ±0.05	0.899 ±0.009	**0.923** ±0.008	0.920 ±0.008
SST-2	0.910 ±0.005	0.920 ±0.008	0.942 ±0.006	**0.945** ±0.008
SentiPolc	0.755 ±0.006	0.760 ±0.008	0.755 ±0.005	**0.770** ±0.005

Results show the mean value of the accuracy on the test set and its standard deviation

(Guarasci et al. 2021, 2024; Gargiulo et al. 2022; Buonaiuto et al. 2024). Given the different nature of the tasks, different evaluation criteria have been used to deepen the qualitative analysis and better understand what positively affects the correct classification and compromises its accuracy. The following analyses were performed on a sample of 100 sentences for each task, balanced by language and dataset.

After the evaluation taking into account linguistic features, a SHAP value analysis has been performed. It aids a qualitative analysis of the results discussed later in this work. It extracts a meaningful explanation about how a quantum transfer learning pipeline can be helpful in described tasks. SHAP (SHapley Additive exPlanations) (Rodríguez-Pérez and Bajorath 2020) is a game-theoretical inspired approach to explain the output of an agnostic parametrized function, such as many deep learning models, using the inputs provided. In essence, SHAP values measure the relative contribution of each feature vector (in the task now considered, a single word embedding) on the model's outcome.

5.3 Syntactic task

Concerning the syntactic task, namely acceptability judgments, intrinsic features of the two datasets have been exploited to perform the qualitative analysis. Both English CoLa and ItaCoLa are provided with fine-grained annotations of linguistic phenomena (Trotta et al. 2021; Warstadt and Bowman 2019). Since each dataset has annotations of different language-dependent phenomena, a shared subset has been chosen. In detail, the phenomena taken into account are as follows:

- **Simple**: sentences with the subject-verb-object structure, in which the subject and arguments are unmodified
- **Binding**: sentences containing bound reflexives or pronouns
- **Question**: sentences with direct or indirect interrogative structure
- **Syntax**: sentences characterized by different syntactic structures (i.e., subject-verb agreement, subordinate and coordinate clauses)

In Table 4, a sample extracted from CoLa of acceptable-unacceptable sentence pairs for each phenomenon is shown. According to the global performances shown in Table 2, it is pretty immediate to see that Bert—both in his classical and quantum versions—is affected by sentence complexity. It, in fact, never fails in simple sentences or in questions whose organization of constituents is respected (hence acceptable). As the complexity of the phenomena increases, with more intricate syntactic structure, indefinite pronouns (*Himself is understood by Rutherford.*), or clause violations (*I know which book Jose didn't read for class, and which book Lilly did it for him.*), the classification is erroneous.

In contrast, the behavior of the models appears different when switching to the Italian language, as shown in Table 5 referring to the ItaCoLa dataset. Contrary to English, simple sentences are not exempt from difficulty in being correctly identified by models. Notice that this category reaches the best results in the original English CoLA corpus (Warstadt and Bowman 2019). This contrast is blamed on the difference between English and Italian word order. English grammar, in fact, is rigorous and forces a strict order. Every English-acceptable sentence presents an SVO (subject-verb-object) order (Liu 2010), making it very easy to process. By contrast, Italian syntax rarely expresses the subject personal pronoun (Chi hai detto... "Who did [you] say...) and is rich in convoluted constructions and ellipses, with an extremely free order of constituents (Brunato et al. 2018), without affecting grammaticality (i.e., Beatrice ha detto che Riccardo crede che Alessandro abbia mentito, "Beatrice said that Richard believes that Alexander lied"). Both Bert and Electra exhibit

Table 4 Overview of linguistic phenomena and predictions by different models both classical and quantum for CoLa dataset in English

Phenomenon	Sentence	Expected result	Bert		Electra	
			Classic	Quantum	Classic	Quantum
Simple	John went home	1	1	1	1	1
	Us love they	0	0	0	0	0
Binding	I talked to Winston about himself	1	0	0	0	1
	Himself is understood by Rutherford	0	1	0	1	0
Question	Where did you go and who ate what?	1	1	1	1	1
	Who does John visit Sally because he likes?	0	1	1	0	0
Syntax	Every senator seems to become more corrupt, as he talks to more lobbyists.	1	1	1	1	1
	I know which book José didn't read for class, and which book Lilly did it for him.	0	1	0	0	1

Table 5 Comparison of classification results on ItaCoLa s using a classic approach based on Bert and Electra and the corresponding quantum transfer learning pipelines

Phenomenon	Sentence	Expected result	Bert		Electra	
			Classic	Quantum	Classic	Quantum
Simple	Alla fine non ha comprato il giornale. (In the end, he didn't buy the newspaper.)	1	0	0	1	0
	I vandali saccheggiarono di Roma. (The Vandals sacked of Rome.)	0	0	0	0	0
Binding	Riccardo ha graffiato se stesso sul viso. (Richard scratched himself on the face.)	1	0	1	1	1
	Alessandro ha laureato se stesso. (Alexander graduated himself.)	0	0	1	0	0
Question	Chi hai detto che credi che Paola pensi che Riccardo abbia incontrato? (Who did you say you believe Paola thinks Riccardo met?)	1	0	0	1	0
	Quali pietre desideri che dormano in non questa stanza? (What stones do you wish would sleep in not this room?)	0	0	0	0	0
Syntax	Beatrice ha detto che Riccardo crede che Alessandro abbia mentito. (Beatrice said that Richard believes that Alexander lied.)	1	0	0	1	0
	Dentro l'armadio, Gabriele ci ha portato. (Inside the closet, Gabriel led us.)	0	1	0	0	1

erratic behavior on phenomena such as binding and syntax. These are pervasive constructions of the Italian language to which pre-trained models have been extensively subjected. Another phenomenon that is critical for Bert and Electra quantum is that of interrogative sentences. Even in this case, this is due to peculiarities of the language, which allows questions with a set of embedded subordinate clauses (i.e., Chi hai detto che credi che Paola pensi che Riccardo abbia incontrato?, "Who did you say you believe Paola thinks Riccardo met?").

A SHAP analysis using the dendrogram formalism has been carried out to better understand which portion of the sentence most affects the correct classification and the differences between the two languages. Dendrograms have undoubted advantages concerning syntactic tasks; they are easily interpreted and are well-suited to comparative analysis. Moreover, they can approximate syntactic relations (Sagae and Gordon 2009).

For instance, concerning simple sentences, the comparison between two unacceptable sentences, "Cynthia chewed," shown on the left in Fig. 5, and "il libro legge" (*the book reads*), shown on the right of the same figure, highlights substantial differences and consequently a different classification.

Notice that in these figures, the discriminating factor is the color of the arches. The ones that positively impact the classification are shown in green, while the ones that undermine it, leading to an erroneous result, are shown in red. As can be clearly seen, there is no problem in classifying the English sentence as unacceptable Fig. 5, since the argument after the verb "chewed," which mandatory requires an object complement, is missing. In the case of the Italian sentence, on the other hand, although it is a short sentence consisting of only three words with no particular lexical difficulty, it is classified incorrectly. This is mainly due to two factors: the construction of the sentence with the subject pronoun "he" omitted (pro-drop) and the left dislocation, with the inversion of the object complement moved to the beginning of the sentence in the preverbal position (il libro, "the book").

The situation is quite different moving to more complex sentences. In this case, sentences shown in Figs. 6 and 7 are syntactically articulated exhibiting different phenomena, such as long-distance relations, subject-verb agreements, or dislocated phrases.

In the sentence extracted from ItaCoLa "E' fuggire che desidera e brama" (*It is running away that Richard desires and yearns.*) shown in Fig. 7, the main clause "[egli] desidera [qualcosa]" (*[he] desires [something]*) is the most relevant portion to classify the sentence as acceptable. However, it is placed after the subordinate clause "E' fuggire che" (*It is running away that*). It is in fact a sentence in which the

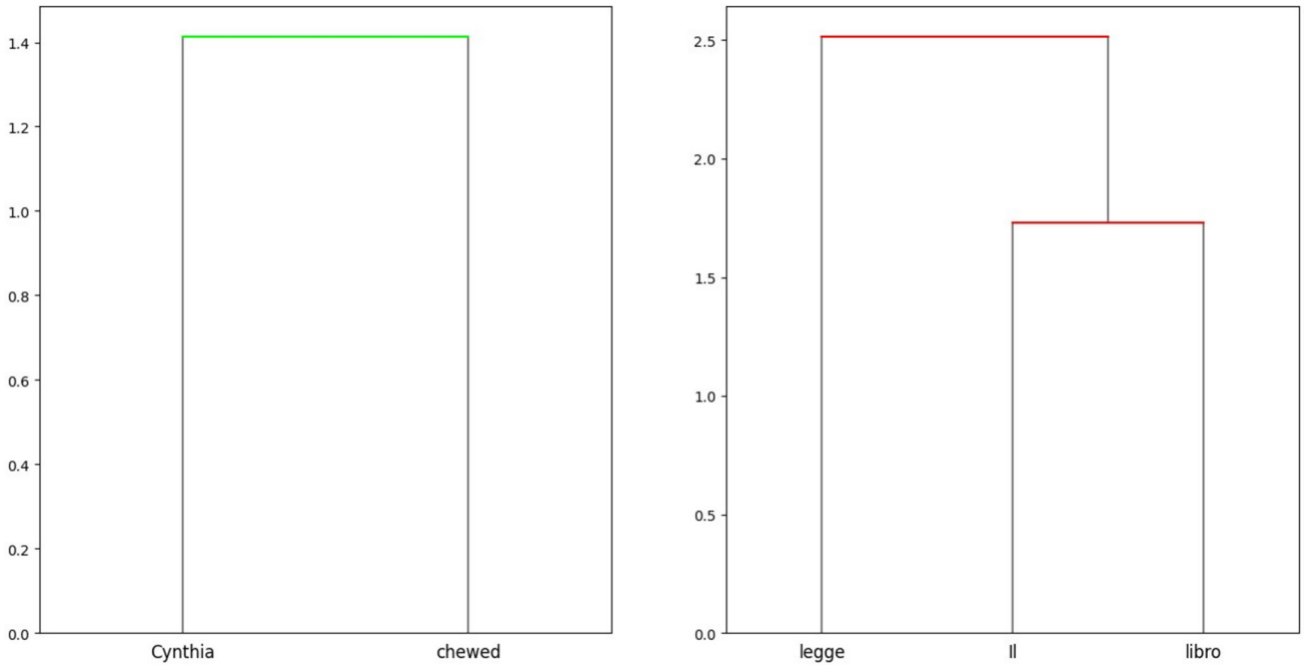

Fig. 5 Example of two simple unacceptable sentences. The first one, in English, is correctly classified (green arcs), while the second, in Italian, is not, as highlighted by red arcs

constituent order is inverted, i.e., the propositional phrase is placed before the main clause instead of after.

The dendrogram of a sentence with similar structure and complexity in English is shown in Fig. 6. Although there is also an inversion in the sentence "It was believed to be illegal by them to do that," it is misclassified as acceptable, despite the fact that it is an ungrammatical sentence. The main clause "to do that" should positively impact the classification, but this does not happen (red arcs), whereas this is the case for the subordinate phrase "It was believed" (green arcs).

Note that an additional factor determining the greater difficulty of the models on the Italian language is due to the

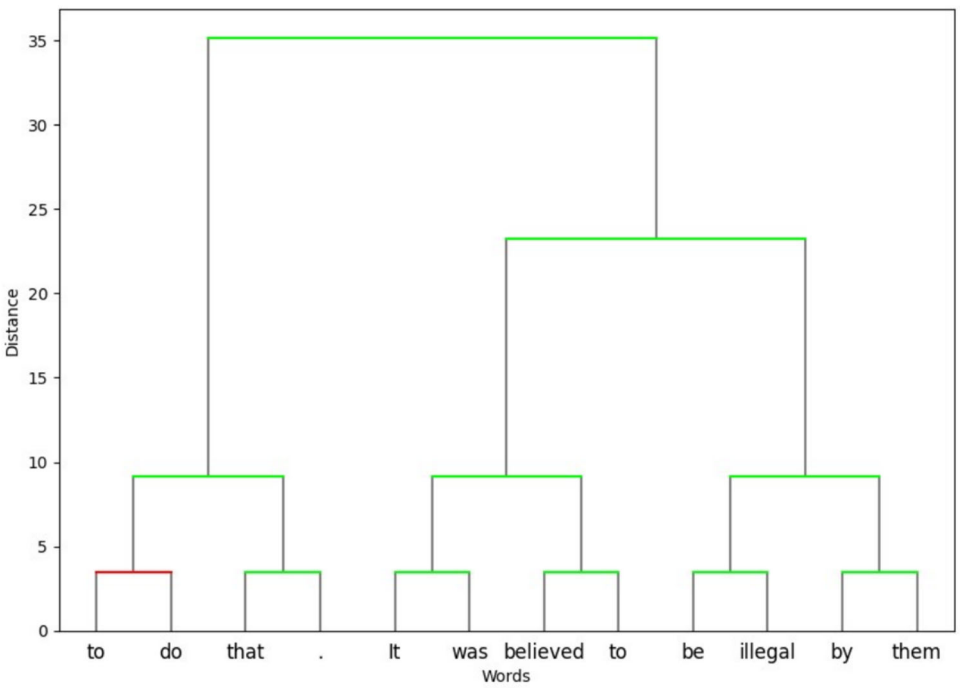

Fig. 6 Example of a dendrogram representation for a simple unacceptable sentence correctly classified in English

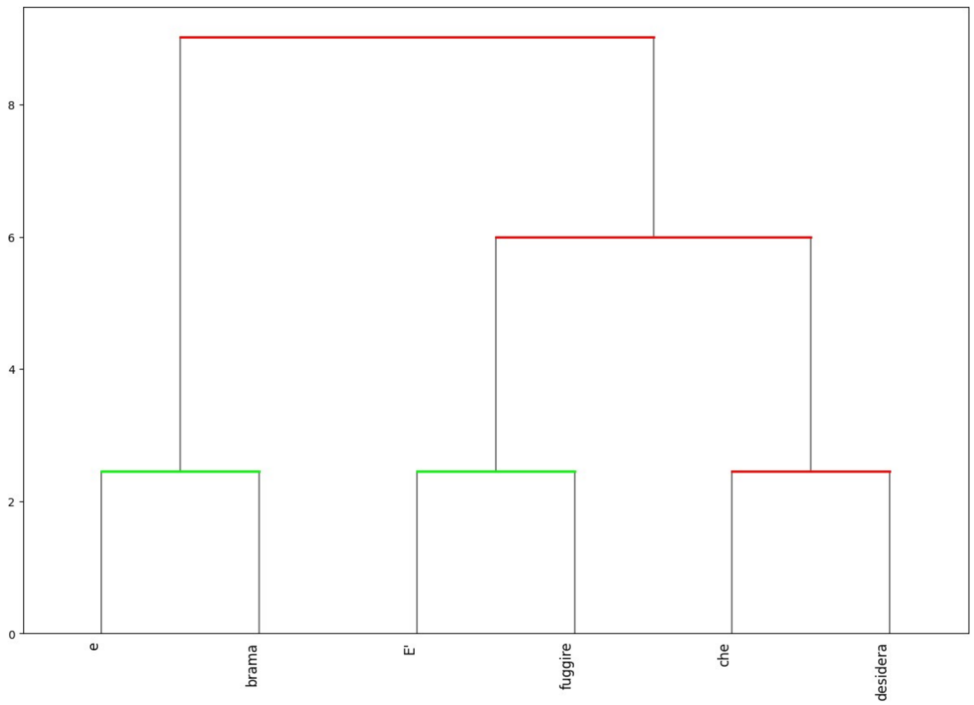

Fig. 7 Example of a dendrogram representation for a complex sentence in Italian. The sentence is unacceptable, but it is misclassified

annotation criterion of ItaCoLa versus that of CoLa. In fact, every sentence in ItaCoLA allows multiple annotations when different phenomena are present. More than 70% of simple sentences also have an additional annotation with at least one other phenomenon, which is definitely a factor in classification errors.

In this case, to explain the misclassification, other elements than just syntactic ones must be considered. Italian sentence is quite complex, containing rare words "brama" (*yearns*), pronoun-dropping (omitted [he]), and the construction with a coordinate following the main sentence, introduced by the conjunction "e" (*and*). However, it is correctly classified. In contrast, what significantly invalidates the correct identification of the unacceptability of the phrase in English is presumably the double subordinating at the beginning of the sentence, in which a cleft construction "It

Table 6 Comparison of the semantic sentiment analysis task results between English and Italian using the different models

Difficulty	Language	Sentence	Expected result	BERT		ELECTRA	
				Classic	Quantum	Classic	Quantum
Difficult (60)	English	sounds like a cruel deception carried out by men of marginal intelligence, with reactionary ideas about women and a total lack of empathy.	0	1	0	0	0
	Italian	@LuzPagoda @bruzziches Pd Pdl Napolitano tuttiacasa M5s bisogna presidiare seggi! Faranno copiosi brogli!In alto gli elemetti E' GUERRA (@LuzPagoda @bruzziches Pd Pdl Napolitano tuttiacasa M5s we must guard polling stations! They will make copious frauds!Up the helmets IT'S WAR!)	0	1	1	1	1
Standard (60–80)	English	the plot is nothing but boilerplate clichés from start to finish	0	0	0	0	0
	Italian	ITALYDESERVESONED Io: No ma grazie eh. (ITALYDESERVESONED Me: No but thanks eh.)	0	1	1	0	0
Easy (80)	English	it 's robert duvall !	1	1	1	1	1
	Italian	Mario Monti nominato Europeo dell'anno (Mario Monti named European of the Year)	1	1	1	1	1

The value 0 indicates a "negative" label, and 1 indicates a "positive" label

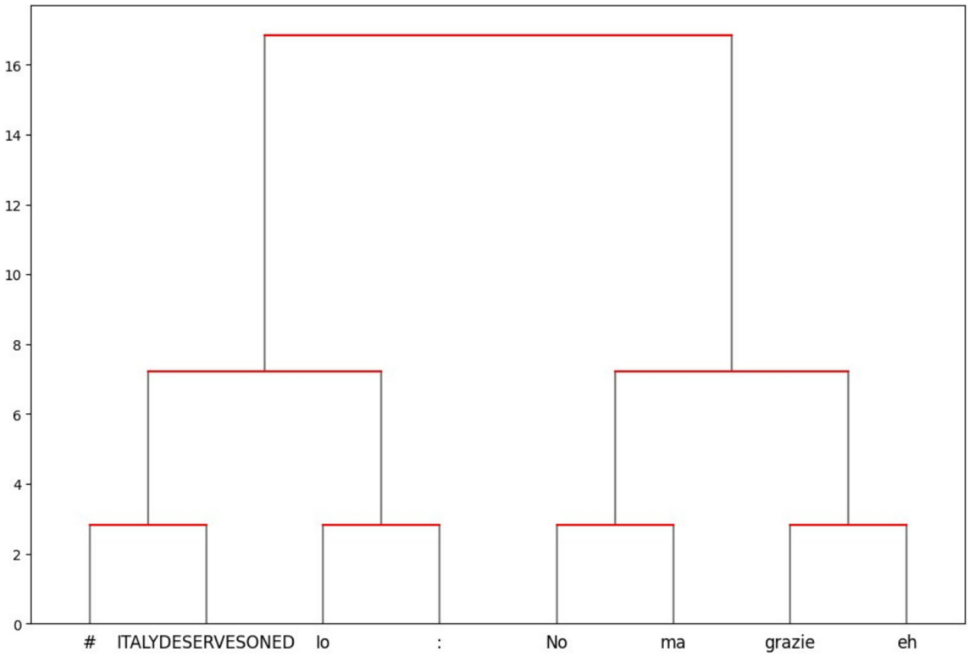

Fig. 8 Example of a dendrogram representation for an incorrectly classified sentence from Sentipolc

was believed" is followed by an infinitive "to be illegal" that then introduces the main proposition displaced to the right at the end of the sentence "to do that."

It pointed out that many factors affect the classification of this syntactic task, not only the judgment (acceptable/unacceptable) but also the lexicon, the complexity of the sentence in terms of subordinate clauses, and the level of nested phrases. Furthermore, another factor can be the annotation criteria of ItaCoLa versus that of CoLa. In fact, every sentence in ItaCoLA allows multiple annotations when different phenomena are present. More than 70% of simple sentences also have an additional annotation with at least one other phenomenon, which is definitely a factor in classification performance.

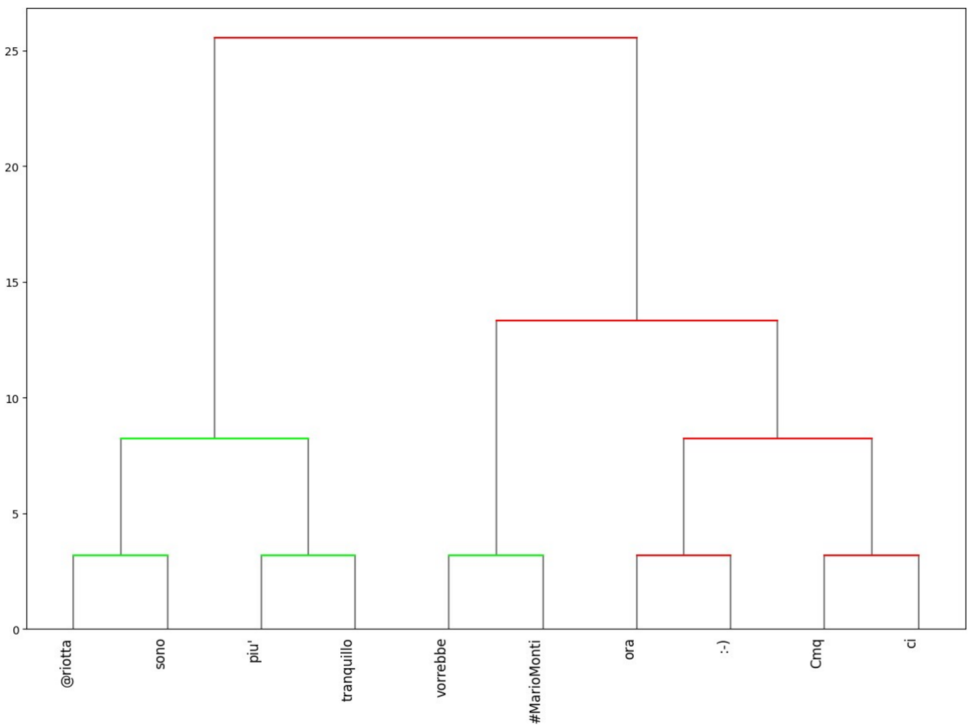

Fig. 9 Example of a dendrogram representation for a correctly classified sentence from Sentipolc

5.4 Semantic task

Different criteria have been adopted to qualitatively evaluate the sentiment analysis task. Although it would have been possible to use the DPTs already present in SST-2, they would have needed to be created ad hoc for SentiPolc, requiring all texts to be subjected to a dependency parser, which could introduce a series of errors. Therefore, the readability index metric has been chosen as the variable to determine sentence complexity.

In this case, since the texts in both datasets are not annotated with criteria that specify their complexity or linguistic properties, the criterion taken into account is grouping by readability score. In this case, since the texts in both datasets are not annotated with criteria that specify their complexity or linguistic properties, the criterion taken into account is grouping by readability score.

Readability is the ease with which a native speaker can understand a text. In the last years, different readability indices to automatically assess a text's quality have been proposed in the literature.

Without going into the debate about the validity or robustness of one index over another, the Flesch-reading-ease test (Eleyan et al. 2020) has been chosen for this work. The reasons are as follows: besides being a widely known index used in many tasks, it is the only one with both versions for English and Italian. Therefore, it can be used for both datasets. Since readability indices use fine-grained groups that are beyond the scope of this paper, sentences have been grouped into three macro classes: challenging to read (Flesch score less than 60), standard and easy to understand (score between 60 and 80), and very simple (score greater than 80).

In Table 6, pair samples from each readability group are shown for English and Italian.

As can be seen immediately, in this case, what most affects the correct classification is not the complexity of the sentence in terms of vocabulary and syntax. Also, no particular terms convey the emotional content, as evidenced in other studies (Guarasci et al. 2024). The aspect that affects this most consistently is data noise.

In fact, sentences that compose SST-2 are extracted from movie reviews, while Sentipolc is entirely composed of comments and posts left by users on Twitter (X), conveying their emotional status. Although the platform has consistently proven to be a valuable resource for sentiment analysis tasks, given its nature as a back-and-forth debate on trending topics with high user engagement, it suffers from significant limitations from a text processing perspective, which have already been highlighted in previous studies (Pota et al. 2020).

Twitter's unique syntax and strict character limit pose significant challenges for NLP. The restriction to 280 characters has motivated users to adopt highly condensed and non-standard language close to spoken jargon. The extensive use of abbreviations ("btw," "u"), acronyms, and special symbols (, #) may affect the immediate comprehension of the text by introducing different degrees of ambiguity. Moreover, the syntax is often frequently alternated with hashtags and mentions, creating a non-conventional grammar.

As evidence of this are the readability scores obtained on Sentipolc sentences in Italian. Even seemingly simple

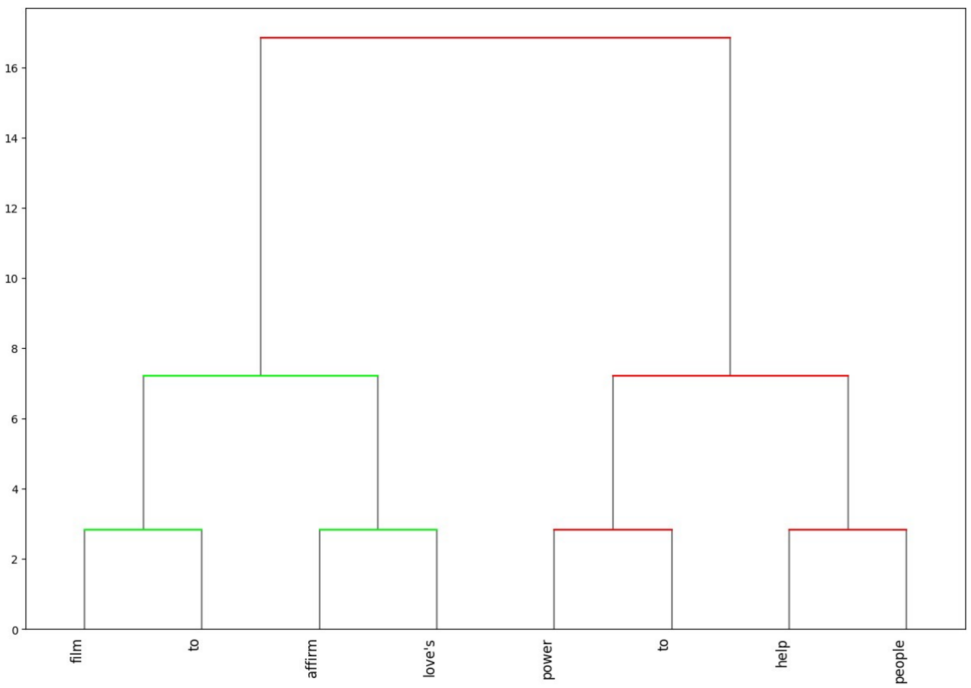

Fig. 10 Example of a dendrogram representation for a correctly classified sentence in English

sentences, such as "#ITALYDESERVESONED Io: No ma grazie eh." (*Me. No but thanks, eh*), which have a concise linear syntax are categorized at the intermediate level (score between 60 and 80) because of the hashtag that opens the sentence. The hashtag #ITALYDESERVESONED compromises the whole syntactic analysis by the models, being a token that is not functional to the semantics of the sentence, as well as an anchor for classification, whose aim is to insert the tweet in a thread with this topic.

Analysis of the dendrograms also shows the same behavior. As shown in Fig. 8, the portion of text that negatively impacts classification is precisely that portion that is apparently taken from the context (#ITALYDESERVESONED). In the following example (Fig. 9), where the hashtag is instead used both to classify the tweet and is inserted within the sentence structure respecting the order of the constituents (#MarioMonti) as a direct object, the classification is carried out correctly, similarly to what happens in a sentence with a comparable level of complexity in English (see Fig. 10).

6 Conclusion

In this study, the potential of quantum transfer learning in NLP has been investigated by evaluating its performance on two distinct tasks: acceptability judgment and sentiment analysis, which are predominantly syntactic and semantic tasks, respectively. Two languages belonging to different families have been taken into account: English and Italian.

Findings indicate that quantum classifiers can achieve competitive performance compared to classical models, particularly in tasks less reliant on syntactic structures. Hence, QTL may offer advantages and have a relevant impact, especially in specific linguistic contexts, due to the unique properties of quantum circuits that allow for more nuanced language representations, although a true quantum advantage is still a subject of open debate in NLP field and beyond (Bravyi et al. 2018 and Zhang et al. 2024).

A significant aspect of the proposed research involved a SHAP-based qualitative analysis, which can provide some insights into the decision-making processes of the models involved in the experiments, both quantum and classic. In particular, features most relevant for the syntactic task are those related to sentence complexity. By contrast, the discriminating element in the semantic task can be mainly attributed to the data source. In this case, there is a bias in the datasets used. Sentipolc, the dataset used for the sentiment analysis in Italian is composed of sentences extracted from Twitter (X), which is very noisy and fragmented. A preliminary data cleaning phase would be needed to allow for better performance.

Concerning the limitations of the proposed approach, the computational demands of quantum models, particularly in terms of time and resources, present challenges for scalability and effective practical deployment. Additionally, as expected, the reliance on classical hardware for quantum circuit evaluation may hinder the realization of quantum computing's full potential in NLP.

Future research should focus on optimizing quantum algorithms to improve efficiency and explore the integration of real quantum hardware to mitigate these limitations. Moreover, a future research perspective is to expand the range of linguistic tasks and languages tested to deepen the understanding of the applicability of quantum transfer learning across diverse contexts. Finally, a more detailed analysis concerning explainability techniques, not only limited to SHAP, may further enrich the depth of future work by offering a more comprehensive overview of the task and techniques used.

Acknowledgements We acknowledge financial support from the project PNR MUR project PE0000013-FAIR.

Author contribution R.G. and G.B. wrote the main manuscript text, G.B. performed the quantum experiments and quantitative analysis, R.G. carried out the qualitative evaluation, M.E. supervised and coordinated the work. All authors reviewed the manuscript.

Funding Open access funding provided by ICAR - NAPOLI within the CRUI-CARE Agreement.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

Al-Qablan TA, Mohd Noor MH, Al-Betar MA, Khader AT (2023) A survey on sentiment analysis and its applications. Neural Comput Appl 35(29):21567–21601

Barnum H, Hayden P, Jozsa R, Winter A (2001) On the reversible extraction of classical information from a quantum source. Proc Royal Soc London. A: Math Phys Eng Sci 457(2012):2019–2039

Basile V, Bolioli A, Bosco C, Nissim M, Patti V, Rosso P, Rabellino S et al (2014) Evalita 2014: Sentipolc Twitter dataset

Basile I, Tamburini F (2017) Towards quantum language models. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 1840–1849

Bel N, Punsola M, Ruíz-Fernández V (2024) Escola: Spanish Corpus of Linguistic Acceptability. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp 6268–6277

Bergholm V, Izaac J, Schuld M, Gogolin C, Ahmed S, Ajith V, Alam MS, Alonso-Linaje G, AkashNarayanan B, Asadi A, Arrazola JM, Azad U, Banning S, Blank C, Bromley TR, Cordier BA, Ceroni J, Delgado A, Di Matteo O, Dusko A, Garg T, Guala D, Hayes A, Hill R, Ijaz A, Isacsson T, Ittah D, Jahangiri S, Jain P, Jiang E, Khandelwal A, Kottmann K, Lang RA, Lee C, Loke T, Lowe A, McKiernan K, Meyer JJ, Montañez-Barrera JA, Moyard R, Niu Z, O'Riordan LJ, Oud S, Panigrahi A, Park C-Y, Polatajko D, Quesada N, Roberts C, Sá N, Schoch I, Shi B, Shu S, Sim S, Singh A, Strandberg I, Soni J, Száva A, Thabet S, Vargas-Hernández RA, Vincent T, Vitucci N, Weber M, Wierichs D, Wiersema R, Willmann M, Wong V, Zhang S, Killoran N (2018) PennyLane: automatic differentiation of hybrid quantum-classical computations. arXiv:1811.04968

Birjali M, Kasri M, Beni-Hssane A (2021) A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl-Based Syst 226:107134

Blank C, Park DK, Rhee J-KK, Petruccione F (2020) Quantum classifier with tailored quantum kernel. npj Quantum Inf 6(1):41

Bonetti F, Leonardelli E, Trotta D, Raffaele G, Tonelli S (2022) Work hard, play hard: collecting acceptability annotations through a 3D game. In: Proceedings of the thirteenth language resources and evaluation conference. European Language Resources Association, pp 1740–1750

Bravyi S, Gosset D, König R (2018) Quantum advantage with shallow circuits. Science 362(6412):308–311

Brunato D, De Mattei L, Dell'Orletta F, Iavarone B, Venturi G (2018) Is this sentence difficult? Do you agree? In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 2690–2699

Buonaiuto G, Guarasci R, Minutolo A, De Pietro G, Esposito M (2024) Quantum transfer learning for acceptability judgements. Quantum Mach Intell 6(1):13

Buonaiuto G, Gargiulo F, De Pietro G, Esposito M, Pota M (2024) The effects of quantum hardware properties on the performances of variational quantum learning algorithms. Quantum Mach Intell 6(1):9

Buonaiuto G, Guarasci R, Esposito M (2024) Quantum transfer learning for sentiment analysis: an experiment on an Italian Corpus. In: Proceedings of the 2024 workshop on quantum search and information retrieval, pp 25–30

Callison A, Chancellor N (2022) Hybrid quantum-classical algorithms in the noisy intermediate-scale quantum era and beyond. Phys Rev A 106(1):010101

Cardillo E, Portaro A, Taverniti M, Lanza C, Guarasci R (2024) Towards the automated population of thesauri using Bert: a use case on the cybersecurity domain. In: International conference on emerging internet, data & web technologies. Springer, pp 100–109

Chen Y, Pan Y, Dong D (2021) Quantum language model with entanglement embedding for question answering. IEEE Trans Cybern

Clark K, Luong M-T, Le QV, Manning CD (2020) ELECTRA: pre-training text encoders as discriminators rather than generators. In: ICLR

Coecke B, de Felice G, Meichanetzidis K, Toumi A (2020) Foundations for near-term quantum natural language processing

Coecke B, Sadrzadeh M, Clark S (2010) Mathematical foundations for a compositional distributional model of meaning. arXiv:1003.4394

Dai A, Xiaohui H, Nie J, Chen J (2022) Learning from word semantics to sentence syntax by graph convolutional networks for aspect-based sentiment analysis. Int J Data Sci Anal 14(1):17–26

Devlin J, Chang M-W, Lee K, Toutanova K. (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota. ACL, pp 4171–4186

Díez-Valle P, Porras D, García-Ripoll JJ (2021) Quantum variational optimization: the role of entanglement and problem hardness. Phys Rev A 104:062426

Eleyan D, Othman A, Eleyan A (2020) Enhancing software comments readability using Flesch reading ease score. Information 11(9):430

Gargiulo F, Minutolo A, Guarasci R, Damiano E, De Pietro G, Fujita H, Esposito M (2022) An Electra-based model for neural coreference resolution. IEEE Access 10:75144–75157

Grant E, Benedetti M, Cao S, Hallam A, Lockhart J, Stojevic V, Green AG, Severini S (2018) Hierarchical quantum classifiers. npj Quantum Inf 4(1):65

Guarasci R, Minutolo A, Damiano E, De Pietro G, Fujita H, Esposito M (2021) Electra for neural coreference resolution in Italian. IEEE Access 9:115643–115654

Guarasci R, De Pietro G, Esposito M (2022) Quantum natural language processing: challenges and opportunities. Appl Sci 12(11):5651

Guarasci R, Silvestri S, De Pietro G, Fujita H, Esposito M (2022) Bert syntactic transfer: a computational experiment on Italian, French and English languages. Comput Speech Lang 71:101261

Guarasci R, Silvestri S, De Pietro G, Fujita H, Esposito M (2022) Bert syntactic transfer: a computational experiment on Italian, French and English languages. Comput Speech Lang 71

Guarasci R, Minutolo A, Buonaiuto G, De Pietro G, Esposito M (2024) Raising the bar on acceptability judgments classification: an experiment on Itacola using Electra. Electronics 13(13):2500

Guarasci R, Catelli R, Esposito M (2024) Classifying deceptive reviews for the cultural heritage domain: a Lexicon-based approach for the Italian language. Expert Syst Appl 252

Guarasci R, Buonaiuto G, De Pietro G, Esposito M (2023) Applying variational quantum classifier on acceptability judgements: a QNLP experiment. In: International conference on numerical computations: theory and algorithms. Springer, pp 98–112

Guarasci R, Silvestri S, De Pietro G, Fujita H, Esposito M (2021) Assessing Bert's ability to learn Italian syntax: a study on null-subject and agreement phenomena. J Ambient Intell Humaniz Comput, pp 1–15

Jentoft M, Samuel D (2023) NoCoLa: the Norwegian corpus of linguistic acceptability. In: Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pp 610–617

Khoo CSG, Johnkhan SB (2018) Lexicon-based sentiment analysis: comparative evaluation of six sentiment lexicons. J Inf Sci 44(4):491–511

Kim J-K, Kim Y-B, Sarikaya R, Fosler-Lussier E (2017) Cross-lingual transfer learning for POS tagging without cross-lingual resources. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 2832–2838

Lambek J (2006) Pregroups and natural language processing. Math Intelligencer 28(2):41–48

Li W, Deng D-L (2024) Extracting reliable quantum outputs for noisy devices. Nat Comput Sci, pp 1–2

Linzen T, Oseki Y (2018) The reliability of acceptability judgments across languages. Glossa: J General Linguistics 3(1)

Liu H (2010) Dependency direction as a means of word-order typology: a method based on dependency treebanks. Lingua 120(6):1567–1578. Contrast as an information-structural notion in grammar

Li Q, Wang B, Zhu Y, Lioma C, Liu Q (2023) Adapting pre-trained language models for quantum natural language processing. arXiv:2302.13812

Li G, Zhao X, Wang X (2022) Quantum self-attention neural networks for text classification. arXiv:2205.05625

Mari A, Bromley TR, Izaac J, Schuld M, Killoran N (2020) Transfer learning in hybrid classical-quantum neural networks. Quantum 4:340

Ma X, Wang Z, Nallapati R, Xiang B (2019) Domain adaptation with Bert-based domain classification and data selection

Mikhailov V, Shamardina T, Ryabinin M, Pestova A, Smurov I, Artemova E (2022) RuCoLa: Russian corpus of linguistic acceptability. arXiv:2210.12814

Möttönen M, Vartiainen JJ, Bergholm V, Salomaa MM (2005) Transformation of quantum states using uniformly controlled rotations. Quantum Info Comput 5(6):467–473

Naresh A, Venkata Krishna P (2021) An efficient approach for sentiment analysis using machine learning algorithm. Evol Intel 14(2):725–731

Pang B, Lee L et al (2008) Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval 2(1–2):1–135

Park DK, Petruccione F, Rhee J-KK (2019) Circuit-based quantum random access memory for classical data. Sci Rep 9(1):3949

Pota M, Ventura M, Catelli R, Esposito M (2020) An effective Bert-based pipeline for twitter sentiment analysis: a case study in Italian. Sensors 21(1):133

Rani S, Kumar P (2019) Deep learning based sentiment analysis using convolution neural network. Arab J Sci Eng 44:3305–3314

Rebentrost P, Schuld M, Wossnig L, Petruccione F, Lloyd S (2019) Quantum gradient descent and newton's method for constrained polynomial optimization. New J Phys 21(7):073023

Ren W, Li W, Xu S, Wang K, Jiang W, Jin F, Zhu X, Chen J, Song Z, Zhang P et al (2022) Experimental quantum adversarial learning with programmable superconducting qubits. Nat Comput Sci 2(11):711–717

Rodríguez-Pérez R, Bajorath J (2020) Interpretation of machine learning models using Shapley values: application to compound potency and multi-target activity predictions. J Comput Aided Mol Des 34(10):1013–1026

Rogers A, Kovaleva O, Rumshisky A (2020) A primer in Bertology: what we know about how Bert works. Trans Assoc Comput Linguist 8:842–866

Ruder S, Ghaffari P, Breslin JG (2017) Knowledge adaptation: teaching to adapt. arXiv:1702.02052

Ruder S, Peters ME, Swayamdipta S, Wolf T (2019) Transfer learning in natural language processing. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: tutorials, pp 15–18

Sagae K, Gordon A (2009) Clustering words by syntactic similarity improves dependency parsing of predicate-argument structures. In: Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09), pp 192–201,

Schuld M, Petruccione F, Schuld M, Petruccione F (2021) Quantum models as kernel methods. Machine Learning with Quantum Computers, pp 217–245

Schuster S, Gupta S, Shah R, Lewis M (2018) Cross-lingual transfer learning for multilingual task oriented dialog. arXiv preprint arXiv:1810.13327

Shah DJ, Lei T, Moschitti A, Romeo S, Nakov P (2018) Adversarial domain adaptation for duplicate question detection. arXiv preprint arXiv:1809.02255

Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 conference on empirical methods in natural language processing, Seattle, Washington, USA. Association for Computational Linguistics, pp 1631–1642

Someya T, Sugimoto Y, Oseki Y (2023) JCoLa: Japanese corpus of linguistic acceptability. arXiv:2309.12676

Stacked Machine Learning (2020) Hybrid model for twitter data sentiment analysis based on ensemble of dictionary based classifier and stacked machine learning classifiers-SVM, KNN and C50. J Theor Appl Inf Technol 98(04):624–635

Sun C, Qiu X, Xu Y, Huang X (2019) How to fine-tune Bert for text classification? In: China national conference on Chinese computational linguistics. Springer, pp 194–206

Tilly J, Chen H, Cao S, Picozzi D, Setia K, Li Y, Grant E, Wossnig L, Rungger I, Booth GH, Tennyson J (2022) The variational quantum eigensolver: a review of methods and best practices. Phys Rep 986:1–128

Trotta D, Guarasci R, Leonardelli E, Tonelli S (2021) Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In: Findings of the association for computational linguistics: EMNLP 2021, Punta Cana, Dominican Republic. Association for Computational Linguistics, pp 2929–2940

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30

Volodina E, Ali Mohammed Y, Klezl J (2021) DaLAJ – a dataset for linguistic acceptability judgments for Swedish. In: Proceedings of the 10th workshop on NLP for computer assisted language learning, Online. LiU Electronic Press, pp 28–37

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S (2018) GLUE: a multi-task benchmark and analysis platform for natural language understanding

Warstadt A, Singh A, Bowman SR (2019) Neural network acceptability judgments. Trans Assoc Comput Linguist 7:625–641

Warstadt A, Bowman SR (2019) Grammatical analysis of pretrained sentence encoders with acceptability judgments. CoRR, abs/1901.03438

Warstadt A, Bowman SR (2019) Linguistic analysis of pretrained sentence encoders with acceptability judgments. arXiv:1901.03438

Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. J Big Data 3:1–40

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Le Scao T, Gugger S, Drame M, Lhoest Q, Rush AM (2019) HuggingFace's transformers: state-of-the-art natural language processing. arXiv e-prints, arXiv:1910.03771

Zaman-Khan H, Naeem M, Guarasci R, Bint-Khalid U, Esposito M, Gargiulo F (2024) Enhancing text classification using bert: a transfer learning approach. Computación y Sistemas 28(4)

Zeng W, Coecke B (2016) Quantum algorithms for compositional natural language processing. Electron Proc Theor Comput Sci 221:67–75

Zhang Z, Gong W, Li W, Deng D-L (2024) Quantum-classical separations in shallow-circuit-based learning with and without noises. Commun Phys 7(1):290

Zhang Z, Liu Y, Huang W, Mao J, Wang R, Hu H (2023) Mela: multilingual evaluation of linguistic acceptability. arXiv:2311.09033