electronics

MDPI

Article

Raising the Bar on Acceptability Judgments Classification: An Experiment on ItaCoLA Using ELECTRA

Raffaele Guarasci [1], Aniello Minutolo [1,*], Giuseppe Buonaiuto [1], Giuseppe De Pietro [2] and Massimo Esposito [1]

[1] Institute for High Performance Computing and Networking (ICAR), National Research Council of Italy (CNR), 80100 Naples, Italy
[2] Department of Information Science and Technology, Pegaso Telematic University, 80143 Naples, Italy
* Correspondence: aniello.minutolo@icar.cnr.it

Abstract: The task of automatically evaluating acceptability judgments has relished increasing success in Natural Language Processing, starting from including the Corpus of Linguistic Acceptability (CoLa) in the GLUE benchmark dataset. CoLa spawned a thread that led to the development of several similar datasets in different languages, broadening the investigation possibilities to many languages other than English. In this study, leveraging the Italian Corpus of Linguistic Acceptability (ItaCoLA), comprising nearly 10,000 sentences with acceptability judgments, we propose a new methodology that utilizes the neural language model ELECTRA. This approach exceeds the scores obtained from current baselines and demonstrates that it can overcome language-specific limitations in dealing with specific phenomena.

Keywords: natural language processing; sentence classification; acceptability judgments; BERT; ELECTRA; low-resource languages

check for updates

Citation: Guarasci, R.; Minutolo, A.; Buonaiuto, G.; De Pietro, G.; Esposito, M. Raising the Bar on Acceptability Judgments Classification: An Experiment on ItaCoLA Using ELECTRA. *Electronics* **2024**, *13*, 2500. https://doi.org/10.3390/electronics13132500

Academic Editors: Leonidas Akritidis and Panayiotis Bozanis

Received: 24 May 2024
Revised: 24 June 2024
Accepted: 25 June 2024
Published: 26 June 2024

1. Introduction

In recent years [1], scholarly interest in acceptability judgments has been rekindled, sparked by the creation of the Corpus of Linguistic Acceptability (COLA) [2], the first large-scale resource collecting acceptability judgments designed specifically to be used for training neural models in the Natural Language Processing field. Such a resource has given rise to a strand of research on this task that started in English and has since expanded to various other languages. Acceptability judgment is a pivotal concept in theoretical linguistics. It can be defined as the assessment of how natural a sentence is perceived by a speaker in his or her native language. Although they are *de facto* recognized as the main source of linguistic data [3,4], there is still a heated debate about the methodologies for collecting and evaluating such judgments [5–7].

Concerning NLP, developing larger and more powerful Neural Language Models (NLMs) has led researchers to explore their capacity to encode various forms of linguistic information. Studies have ranged from investigating specific linguistic phenomena to general grammar knowledge [8–12]. In this context, acceptability judgments have emerged as a crucial domain for evaluating the linguistic knowledge acquisition of these models [13,14], mainly since COLA has been incorporated into the widely used GLUE evaluation benchmark [2]. Subsequently, similar resources have been released in different languages, including those belonging to very different language families: Russian [15], Japanese [16], Norwegian [17], Swedish [18], Spanish [19], and Italian [20], which is the language that is the subject of this work.

These datasets have been typically used with monolingual and cross-lingual approaches to assess the syntactic abilities of NLMs or to evaluate the goodness of models in natural language generation tasks [21]. Currently, BERT-based models are the ones that achieve the best performance.

This paper proposes an approach for the Italian language using ELECTRA [22] for the acceptability task, demonstrating that it can exceed the performance currently achieved in the literature. In recent years, ELECTRA has demonstrated that it can overcome BERT in different NLP tasks [23,24] with equal size and available resources, with a particular focus on its application in languages other than English [25,26]. The dataset on which the model is tested is ItaCoLA [20], the largest current resource in the Italian language for acceptability judgments. ItaCoLA includes around 9700 sentences from linguistic scientific literature spanning four decades. These sentences have been manually transcribed and converted into digital format [27].

Adhering to the prevailing methodology in this domain, the acceptability assessment is based on binary judgments, as determined by expert linguists. In addition to the quantitative analysis, which compares performance with the current ItaCoLA baseline, a qualitative analysis is also proposed, which takes advantage of the large number of linguistic phenomena covered by the corpus and the manual annotation [28,29].

Notice that the work's primary contribution is to demonstrate how applying a model like ELECTRA, whose main feature is to achieve superior performance to models such as BERT with lower computational cost and fewer examples, can improve performance on acceptability judgment tasks in the Italian language. Although the task of predicting acceptability judgments has been heavily discussed in recent years, and more resources are being released in other languages, often the time costs and complexity of the models used make it difficult to achieve a very satisfactory cost–benefit ratio [30], even if results are promising.

The paper's organization is the following: Section 2 briefly describes recent works on acceptability judgments. Section 3 describes the resources and models taken into account, along with the experiment setup. In Section 4, the results of the analyses, both from a quantitative and qualitative perspective, are presented and discussed. Finally, Section 5 summarizes the work and provides the conclusions.

2. Related Work

Accurately classifying acceptability judgments has always been a popular topic of discussion in linguistics because of its theoretical aspects related to cognitive science or issues concerning the connection between syntax and knowledge [31]. Concerning NLP, the task is at the heart of many applications, ranging from simple tasks such as grammar correction to more elaborate ones such as machine translation and evaluation of automated dialogue systems. Consequently, several challenges arise in this context. The first issue is the subjective nature of acceptability judgment, which can vary according to context and language and is influenced by syntactic or semantic features as well as pragmatic and dialogic ones. Therefore, models facing this task must be able to identify and capture complex linguistic structures [32] and exploit cross-lingual approaches to generalize between languages [33].

Moreover, an additional bottleneck is the cost and difficulty of obtaining annotated data that may suit these models. Often, it is necessary to rely on crowdsourcing or the support of domain experts.

The event that caused great traction for the assessment of acceptability tasks has been the public release of the CoLa corpus [2], the most extensive existing English acceptability corpus that includes over 10,000 sentences. Numerous neural network-based approaches were compared on the CoLA corpus, which was then incorporated into the widely known natural language understanding (NLU) benchmark dataset GLUE [34].

Regrettably, most studies within GLUE have reported accuracy instead of the Matthews Correlation Coefficient (MCC), making it challenging to determine the optimal approach. Nevertheless, it is noteworthy that top-ranking systems are transformer-based models, i.e., ALBERT [35] (69.1 Accuracy), and StructBERT [36] (69.2 Accuracy). Instead, another line of research has approached the task using entailment and exploiting small-scale models [37] showing promising results (86.4 Accuracy).

The methodology introduced in CoLA has been the starting point for several derivative resources developed recently and focused on languages other than English. Such languages include Italian [20], Norwegian [17], Swedish [18], Russian [15], Japanese [38], Chinese [39,40], and Spanish [19]. It is important to note, however, that since acceptability has always fascinated scholars, small datasets had already been released before CoLa, mainly focused on theoretical linguistics or cognitive science-related tasks [41–43]. In addition to English, informal acceptability judgments have been evaluated in Hebrew and Japanese [32], as well as in French [44] and Chinese [45]. A small Italian dataset focusing on complexity and acceptability has also been released [46]. Notice that—in the context of the newborn field of Quantum Natural Language Processing (QNLP)—ItaCola has been used to evaluate the feasibility of a quantum machine learning algorithm to classify acceptable/unacceptable sentences using the new distributional compositional models of language [47].

3. Materials and Methods

3.1. Dataset

The resource employed in this work is ItaCoLA, which stands for the Italian Corpus of Linguistic Acceptability [20]. ItaCoLA has been meticulously constructed to encompass a diverse spectrum of linguistic phenomena while making a clear distinction between sentences regarded as acceptable and those deemed unacceptable. The process used to curate this corpus has been closely modeled after the methodology applied in creating the original CoLA [2].

ItaCoLA consists of 9700 sentences whose origins vary. These sentences encompass a wide array of linguistic phenomena for comprehensive coverage of the linguistic literature. The acceptability assessment of each sentence comes from experts who authored the diverse data sources and is formulated as a binary score.

The sentences have been collected from a wide range of linguistic publications spanning four decades, meticulously transcribed by hand, and made available in digital format. A sample extracted from ItaCoLA with some acceptable sentences (label 1) and some unacceptable ones (label 0) is shown in Table 1.

As mentioned above, the annotation process lies in domain-expert judgments. This procedure, already known in corpus linguistics studies [48], has become the standard de facto for this type of task, shared by all the works in other languages derived from CoLa. The possibility of using crowdsourcing approaches and naive annotators is still debated in the literature [49], as well as creating deliberately unacceptable examples ad-hoc by compromising well-formed sentences [50], a procedure widely used in other NLP tasks, such as sentiment analysis or fake news detection [51–54].

Table 1. Sentences from ItaCoLA. The first column indicates the acceptability judgment (1 = acceptable, 0 = not acceptable).

Label	Sentence
0	Maria andava nella sua l'inverno passato città. (Maria went to her winter past city)
1	Max vuole sposare Andrea (Max want to marry Andrea)
0	Il racconto ti hanno colpito. (The story have impressed you)
1	Il racconto ti ha colpito. (The story has impressed you)

ItaCoLA is divided as follows: 7801 sentences compose the test set, the validation set includes 946 sentences, while the test set is 975. The ratio of acceptable to unacceptable sentences in each split is balanced.

3.2. Models

3.2.1. BERT

In the realm of NLMs, BERT has emerged in the literature as the most widely adopted model due to its remarkable efficiency [55]. BERT is based on a Transformer encoder [56], and it needs several non-annotated data for the training phase, articulated in two different training objectives, namely masked language modeling (MLM) and next sentence prediction.

MLM entails randomly masking a portion of words of the training dataset. This technique enables the model to capture information bidirectionally within sentences while simultaneously predicting the masked words. It is worth noting that two possible options for vocabulary (cased or uncased) imply two distinct pre-trained models. This bidirectional analytical adaptability allows the model to maintain a significant generative capacity through the inner layers of the network while also facilitating adaptation to specific tasks during the subsequent fine-tuning phase.

BERT operates by initiating each input word sequence with a special token, marked as "*[CLS]*". This token is crucial in deriving an output vector of size H, corresponding to the hidden layers' dimensions and the whole input sequence. Furthermore, another unique token, "*[SEP]*", needs to be correctly situated within the input sequence following each sentence. Starting from a sequence of input words denoted as $t = (t_1, t_2, \ldots, t_m)$, BERT produces an output represented as $h = (h_0, h_1, h_2, \ldots, h_m)$. In this representation, $h_0 \in \mathbb{R}^H$ is the ultimate hidden state of the special token "*[CLS]*", acting as a comprehensive representation for the entire input sequence. Meanwhile, $h_1, h_2, \ldots, h_m$ signify the final hidden states of the remaining input tokens.

The context-dependent representation of sentences obtained from this training phase can be further customized to specific tasks by fine-tuning and modifying several hyperparameters. The $BERT_{base}$ model has 110 million parameters (12 hidden layers, each composed of 768-dimensional states and 12 attention heads). Every layer of the model produces a unique embedded representation of the input words, whose dimension is limited to a maximum of 512 tokens.

For the fine-tuning of BERT in classifying input sequences of words into K distinct text categories, the final hidden state h_0 can serve as the input to a classification layer. Subsequently, a softmax operation [57] is employed to transform the scores corresponding to each text category into probabilities.

$$P = softmax(CW^T) \tag{1}$$

The parameter matrix of the classification layer as $W \in \mathbb{R}^{K \times H}$ is the one selected for this work. Concerning the BERT version, the *dbmdz* Italian BERT model (XXL, cased) [58] has been chosen. It is an Italian pre-trained version of BERT trained using different corpora [59,60]. The corpus used for the training is 81 GB and includes 13,138,379,147 tokens.

3.2.2. ELECTRA

The other NLM under consideration is ELECTRA, first introduced in [22]. ELECTRA has demonstrated superior proficiency in capturing contextual word representations, surpassing other models in downstream performance when subjected to identical model sizes, data, and computational resources, as noted by [61]. ELECTRA's pre-training includes two transformer models: the generator (G) and the discriminator (D), as shown in Figure 1. G is devoted to replacing some tokens in a sequence, typically trained as a masked language model. In contrast, the main focus in ELECTRA is on the discriminator model D, which aims to discern the tokens substituted by G in the sequence.

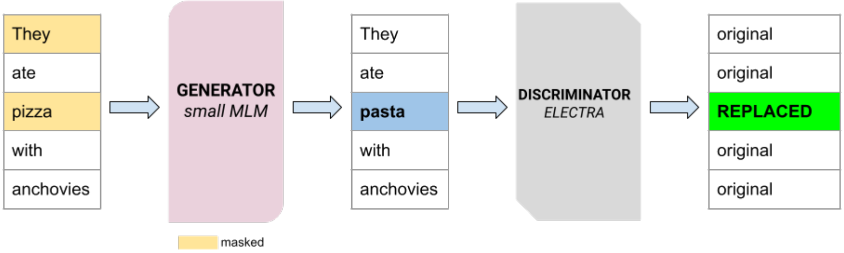

Figure 1. ELECTRA overview with replaced token detection.

In a specific scenario, when certain tokens within a given input sequence are randomly substituted with a unique "*[MASK]*" token, the aim of G is to predict the original tokens for all the masked instances. Following this, a sequence with fake tokens is generated for D, which is trained to distinguish genuine from fake tokens using a method called replaced token detection (RTD). The RTD offers the advantage of not compromising the model's overall performance while having fewer examples available.

Similarly to BERT, a version of the model used is *dbmdz* Italian ELECTRA [62]. Here are the details: starting from a sentence κ of raw text χ, made of a set of tokens $\kappa = w_1, w_2, \ldots, w_n$ where w_t $(1 \leq t \leq n)$ is a generic token, κ is encoded in a sequence of contextualized vector representations $h(\kappa) = h_1, h_2, \ldots, h_n$ by G and D. After that, using a softmax layer, G is the probability of generating a specific token w_t for each position t for which $w_t = [MASK]$.

$$p_G(w_t|\kappa) = \frac{r(w_t)^T h_G(\kappa)_t}{\sum_{w'} exp(r(w')^T h_G(\kappa)_t)} \tag{2}$$

The embedding function is represented by $r(\cdot) : w_t \in \kappa \to \mathbb{R}^{dim}$; dim is the chosen embedding size, while the prediction of whether w_t is original or fake is given by D. A sigmoid layer, σ, is used to perform this task:

$$D(\kappa, t) = \sigma(r(w_t)^T h_D(\kappa)_t) \tag{3}$$

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

$$\min_{\eta_G, \eta_D} \sum_{\kappa \in \chi} \mathcal{L}_{Gen}(\kappa, \eta_G) + \lambda \mathcal{L}_{Dis}(\kappa, \eta_D) \tag{4}$$

Note that $\mathcal{L}_{\text{Gen}}$ represents the loss function of G and $\mathcal{L}_{\text{Dis}}$ that of D. Subsequently, only D is used for the fine-tuning.

Techniques like MLM, exemplified by BERT, introduce input corruption by substituting a masked token for an original one, which the trained model then retrieves. Such methods yield commendable results when applied to downstream NLP tasks; they typically demand substantial computational resources for optimal effectiveness.

By contrast, RTD provides a more efficient pre-training technique, corrupting a subset of input tokens with plausible alternatives using a generator network. ELECTRA's efficiency compared to models such as BERT lies in including the predictions of all input tokens, not only the masked ones. Therefore, D loss can also be computed on the whole set of tokens in the input sequence, allowing the use of examples in the training phase without compromising performance.

4. Results and Discussion

Two different types of analysis have been carried out: quantitative, which aims to verify the performance achieved by the tested models using metrics well known in the literature and the improvement from the baseline, and qualitative, which aims to deepen the analysis and estimate whether there are specific phenomena that impact performance the most.

4.1. Quantitative Analysis

According to previous studies approaching this task, two different metrics have been used for the analysis: accuracy and the Matthews Correlation Coefficient (MCC). Accuracy is the most commonly used basic metric and is also the one used to be able to compare with GLUE. MCC is a correlation metric increasingly used in binary classification tasks [63].

The Adam optimizer has been used for training (learning rate of 2×10^{-5}, epsilon of 10^{-8}), while batch size has been set to 32, with 2 labels, 0 warm-up steps, a maximum input sequence length of 64 words, categorical cross-entropy as the objective function. The number of epochs on which the model has been trained is 7.

As evidenced by the loss functions shown in Figure 2, ELECTRA is more efficient than BERT at loss-minimizing learning to perform classification.

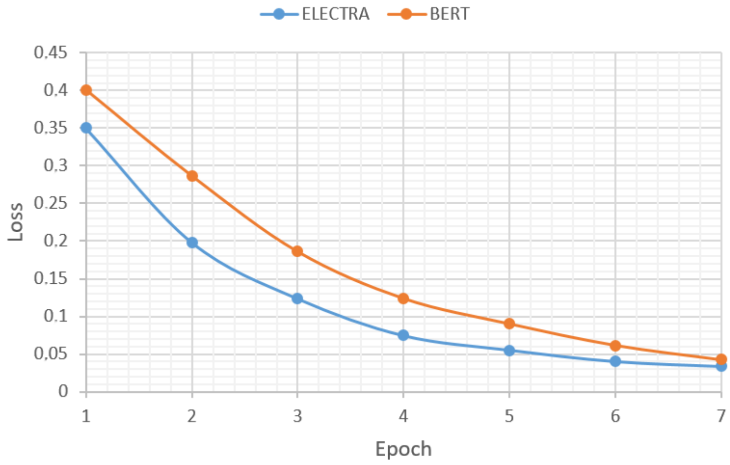

Figure 2. The loss functions for BERT and ELECTRA training.

Concerning NLMs tested, as shown in Table 2, it can be noted that both BERT and ELECTRA outperform the classic LSTM baseline.

Table 2. Classification results comparing LSTM, BERT, and ELECTRA.

Model	Accuracy	MCC
LSTM	0.794	0.278 ± 0.029
BERT	0.904	0.603 ± 0.022
ELECTRA	0.923 ± 0.008	0.690 ± 0.035

Notice that, although it is an outdated architecture [64], LSTM (Long Short-Term Memory) models have been used in several works for acceptability classification [2,65]. Their peculiarity is, in fact, their ability to adequately capture and handle long-term dependencies in sequential data. Furthermore, LSTM memory cells are able to maintain long-term information, unlike traditional RNNs.

In particular, experiments carried out using ELECTRA achieve the best results, reaching an accuracy of 0.923, while the BERT-Classic reaches a lower score, ending at 0.904. By using MCC as a metric, the result is even more significant.

This result is attributable to a more efficient use of available data. ELECTRA's pre-training strategy is not limited to learning only masked words, as with BERT. Unlike MLM, RTD produces a better contextual representation by learning from all input words and using a similar amount of data, model size, and computational cost [66]. Further confirmation of the validity of this approach is given by its application in similar binary classification tasks in different languages [53,67,68].

4.2. Qualitative Analysis

The evaluation has also been extended to a qualitative level to take advantage of fine-grained annotations provided along with ItaCoLA.

Since its release, around 30% of the sentences composing the corpus have been annotated using labels covering nine linguistic phenomena, as shown in Table 3. The phenomena combine some classes proposed for the AcComplit dataset [46] and other ones used in [69] for the English language.

Table 3. Overview of different phenomena collected in ItaCola.

Phenomenon	Sentences	Description	Example
Simple	365	One-verb sentences composed of only mandatory arguments.	"Marco ha baciato Alice" (En. *Marco kissed Alice.*)
Cleft constructions	136	Sentences in which a constituent is displaced from its typical position to give it emphasis.	"È Clara che Anna ha visto uscire" (En. *It is Clara whom Anna saw leaving*.)
Subject–verb agreement	406	Sentences lacking the agreement in gender or number between subject and verb.	"Maurizio sostiene che Lucia ha parlato di lui a casa con la moglie" (En. *Maurizio claims that Lucia talked about him at home with his wife.*)
Indefinite pronouns	312	Sentences with one or more indefinite pronouns referring to someone or something.	"Spero in qualcosa che arriverà" (En. *I am hoping for something to come.*)
Copular constructions	855	Sentences in which the subject is connected to a noun or an adjective with a copulative verb.	"Cicerone era un grande oratore" (En. *Cicero was a great speaker*.)
Auxiliary	398	Sentences containing the verb "essere" (to be) or "avere" (to have).	"Stavamo correndo nel pomeriggio" (En. *We were running in the afternoon.*)
Bind	27	Sentences in which anaphoric elements are grammatically associated with their antecedents.	"Cesare adula se stesso" (En. *Caesar flatters himself.*)
Wh-islands violations	53	Sentences at the beginning of which there is a Wh- clause.	"Che opera lirica avevi suggerito di andare a vedere stasera?" (En. *What opera did you suggest we see tonight?*)
Questions	177	Interrogative sentences.	"È tua quella bicicletta rossa?" (En. *Is that red bicycle yours?*)

Since only 2088 sentences are accompanied by a fine-grained linguistic annotation, the train, test, and validation splits have been altered to achieve this objective: the whole set comprising all the 2088 sentences is designated the test set. Therefore, the remaining 7632 sentences in the dataset have been divided into two subsets, training and validation, which are composed of 6833 and 800 sentences, respectively. ELECTRA has undergone fine-tuning using identical parameters to those in previous experiments.

Concerning accuracy, as reported in Figure 3, some uniformity can be seen with a significant gap only in sentences belonging to the bind class.

As expected, sentences involving pervasive constructions of the Italian language are simpler for the model to handle. This is true for copular constructions and questions. Such sentences achieve almost identical results (accuracy equal to 0.88 and 0.86, respectively). In the other phenomena, on the other hand, the deviation is very low, in the range of 3 points.

The only exception, as mentioned above, concerns the bind class, classified poorly using BERT (0.55) but which undergoes a significant increase using ELECTRA (0.70). This is a very interesting result since binding is a complex phenomenon studied in various languages, related to well-known concepts in theoretical linguistics such as anaphora and ergative verbs [70,71], which have often posed numerous critical issues in NLP [72].

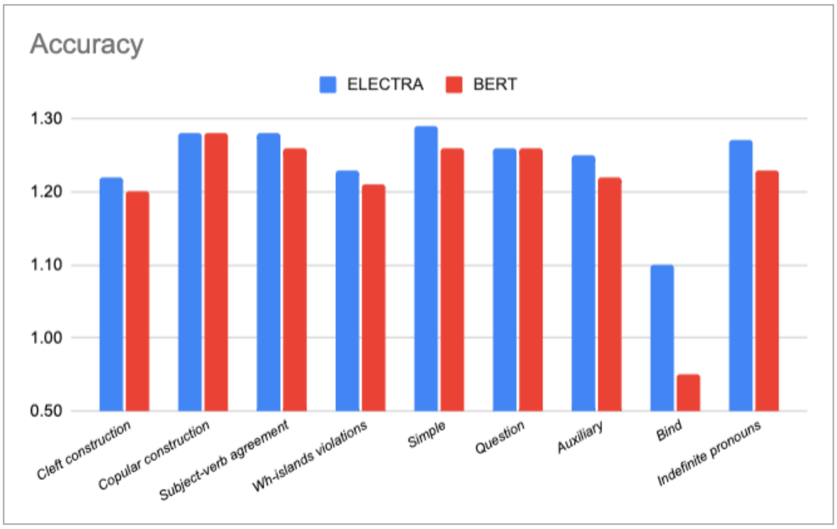

Figure 3. Comparison of performance in terms of accuracy between BERT and ELECTRA.

Unexpectedly, simple sentences do not yield the highest results. In contrast, this category achieves the best outcomes in the original English CoLA corpus [69]. This discrepancy can be attributed to English's extremely straightforward syntax and strict SVO (subject–verb–object) order [65], factors that contribute to sentences lacking any particular criticality.

The same is not true for Italian, where the syntax is often characterized by hypotaxis and the presence of pro-drop, and the free order of constituents constitute major critical factors that affect performance [73].

Furthermore, it is notable that ItaCoLA allows multiple annotations for the same sentence in case phenomena coexist.

Almost a third of the sentences in the dataset used for this experiment have multiple annotations. As for simple sentences, 77% have more than one annotation, which could be another reason they tend to be misclassified.

Overall, as shown in Table 4, the application of ELECTRA achieves values consistently better than or equal to BERT, both using accuracy and MCC in every phenomenon.

Table 4. Results of two models using MCC and Accuracy (ACC) with respect to each phenomenon taken into account.

Phenomenon	Model	
	ELECTRA	BERT
	MCC / ACC	
Cleft construction	0.53/0.82	0.48/0.80
Copular construction	0.56/0.88	0.36/0.88
Subject–verb agreement	0.54/0.88	0.41/0.86
Wh-islands violations	0.5 /0.83	0.46/0.81
Simple	0.54/0.89	0.35/0.86
Question	0.50/0.86	0.37/0.86
Auxiliary	0.47/0.85	0.30/0.82
Bind	0.43/0.70	0.18/0.55
Indefinite pronouns	0.51/0.87	0.28/0.83
Total	0.54/0.87	0.37/0.84

Considering the MCC as a metric, a major variability across phenomena can be observed (see Figure 4). An issue highlighted at the release of ItaCoLA was the low performance on the copula constructions and Wh-violations. This result strongly contrasted with the results obtained for English: in [69], a value of MCC > 0.50 was presented for both phenomena. This problem seems to be overcome using ELECTRA; in both cases, the values sharply increase, reaching MCC scores of 0.56 and 0.58, which is in line with English CoLa scores. Although interesting from a cross-linguistic perspective, it should be noted that many of these phenomena are highly language-specific. Therefore, a true Italian–English comparison for each phenomenon is not possible.

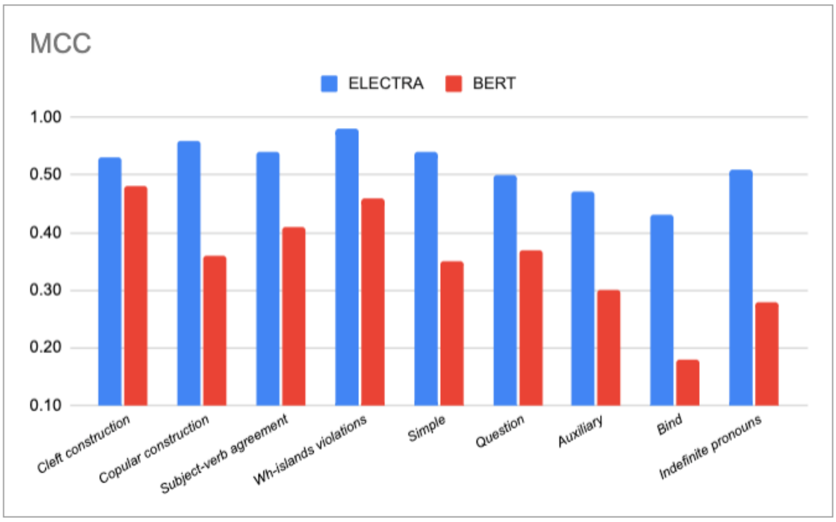

Figure 4. Comparison of performance in terms of MCC between BERT and ELECTRA.

5. Conclusions and Future Work

In this work, an approach that raises the bar for the performance of acceptability judgment tasks in Italian has been presented. In particular, using the ELECTRA model has enabled surpassing baselines and state-of-the-art BERT-based approaches.

ELECTRA performance has also been investigated in depth through a qualitative analysis that focused on specific linguistic phenomena, showing a generalized improvement, particularly regarding marginal phenomena poorly represented in the sample under analysis, in which BERT has been underperforming.

Following the insight already presented in [15], the work's future development consists of exploring the possibilities of cross-lingual approaches [74].

Obviously, many open issues cannot disregard the nature of the task itself, since the unacceptability of certain syntactic structures is strictly language-dependent. For this reason, it would be fruitless to compare global performance through cross-linguistic approaches; rather, it would be appropriate to focus on specific phenomena, as already demonstrated in other studies in the literature [75]. Concerning further experiments, additional models, such as decoder-only or encoder–decoder models, will be tested, and the effect of in-context learning and knowledge transfer from additional languages will be considered, following the most recent research trends in this topic.

Finally, given the recent interest in the syntactic evaluation of NLMs, to make the methodology more robust, a comparison with experienced and unskilled human annotators will be introduced, as proposed in [28,49], and a semi-automatic systematic evaluation system based on a set of minimal pairs, as has happened with English [76] and Japanese [38]. Furthermore, new lines of research will be investigated concerning the promising results in the area of QNLP obtained from the preliminary experiments [77] and the chance to also opt for different strategies based on zero or few-shot learning using other NLMs on this task [78].

Author Contributions: Conceptualization, R.G., A.M. and M.E.; Methodology, R.G., A.M., G.D.P. and M.E.; Software, A.M.; Validation, R.G.; Formal analysis, G.B.; Resources, R.G.; Writing—original draft, R.G. and A.M.; Writing—review & editing, G.B., G.D.P. and M.E.; Supervision, M.E.; Funding acquisition, M.E. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge financial support from the H2IOSC Project—Humanities and Cultural Heritage Italian Open Science Cloud funded by the European Union—NextGenerationEU—NRRP M4C2—Project code IR0000029-CUP B63C22000730005.

Data Availability Statement: The data presented in this study are openly available in GitHub. [https://github.com/dhfbk/ItaCoLA-dataset] (accessed on 24 April 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, S.Y.C.; Huang, C.M.; Hsing, C.W.; Kao, Y.J. Hybrid quantum-classical classifier based on tensor network and variational quantum circuit. *arXiv* **2020**, arXiv:2011.14651.
2. Warstadt, A.; Singh, A.; Bowman, S.R. Neural Network Acceptability Judgments. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 625–641. [CrossRef]
3. Chomsky, N. *Aspects of the Theory of Syntax*; MIT Press: New York, NY, USA, 1965.
4. Schütze, C.T. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*; University of Chicago Press: Chicago, IL, USA, 2016.
5. Gibson, E.; Fedorenko, E. The need for quantitative methods in syntax and semantics research. *Lang. Cogn. Process.* **2013**, *28*, 88–124. [CrossRef]
6. Sprouse, J.; Almeida, D. A quantitative defense of linguistic methodology. 2010, *Manuscript Submitted for Publication*.
7. Linzen, T. What can linguistics and deep learning contribute to each other? Response to Pater. *Language* **2019**, *95*, e99–e108. [CrossRef]
8. Hewitt, J.; Manning, C.D. A Structural Probe for Finding Syntax in Word Representations. 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, MN, USA, 2–7 June 2019; pp. 4129–4138. [CrossRef]
9. Manning, C.D.; Clark, K.; Hewitt, J.; Khandelwal, U.; Levy, O. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 30046–30054. [CrossRef] [PubMed]
10. Jawahar, G.; Sagot, B.; Seddah, D. What does BERT learn about the structure of language? In Proceedings of the ACL 2019-57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
11. Guarasci, R.; Damiano, E.; Minutolo, A.; Esposito, M.; De Pietro, G. Lexicon-grammar based open information extraction from natural language sentences in Italian. *Expert Syst. Appl.* **2020**, *143*, 112954. [CrossRef]
12. Esposito, M.; Damiano, E.; Minutolo, A.; De Pietro, G.; Fujita, H. Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. *Inf. Sci.* **2020**, *514*, 88–105. [CrossRef]
13. Gulordava, K.; Bojanowski, P.; Grave, É.; Linzen, T.; Baroni, M. Colorless Green Recurrent Networks Dream Hierarchically. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; pp. 1195–1205.
14. Lau, J.H.; Armendariz, C.; Lappin, S.; Purver, M.; Shu, C. How Furiously Can Colorless Green Ideas Sleep? Sentence Acceptability in Context. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 296–310. [CrossRef]
15. Mikhailov, V.; Shamardina, T.; Ryabinin, M.; Pestova, A.; Smurov, I.; Artemova, E. RuCoLA: Russian Corpus of Linguistic Acceptability. *arXiv* **2022**, arXiv:2210.12814.
16. Someya, T.; Sugimoto, Y.; Oseki, Y. JCoLA: Japanese Corpus of Linguistic Acceptability. *arXiv* **2023**, arXiv:2309.12676.
17. Jentoft, M.; Samuel, D. NocoLA: The norwegian corpus of linguistic acceptability. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), Tórshavn, Faroe Islands, 22–24 May 2023; pp. 610–617.
18. Volodina, E.; Mohammed, Y.A.; Klezl, J. DaLAJ—A dataset for linguistic acceptability judgments for Swedish. In Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning, Online, May 2021; pp. 28–37.
19. Bel, N.; Punsola, M.; Ruíz-Fernández, V. EsCoLA: Spanish Corpus of Linguistic Acceptability. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, 20–25 May 2024; pp. 6268–6277.
20. Trotta, D.; Guarasci, R.; Leonardelli, E.; Tonelli, S. Monolingual and Cross-Lingual Acceptability Judgments with the Italian CoLA corpus. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 16–20 November 2021; pp. 2929–2940.
21. Volodina, E.; Mohammed, Y.A.; Berdičevskis, A.; Bouma, G.; Öhman, J. DaLAJ-GED-a dataset for Grammatical Error Detection tasks on Swedish. In Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning, Online, May 2023; pp. 94–101.
22. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In Proceedings of the ICLR, Addis Ababa, Ethiopia, 26–30 April 2020.
23. Fang, H.; Xu, G.; Long, Y.; Tang, W. An Effective ELECTRA-Based Pipeline for Sentiment Analysis of Tourist Attraction Reviews. *Appl. Sci.* **2022**, *12*, 10881. [CrossRef]
24. Gargiulo, F.; Minutolo, A.; Guarasci, R.; Damiano, E.; De Pietro, G.; Fujita, H.; Esposito, M. An ELECTRA-Based Model for Neural Coreference Resolution. *IEEE Access* **2022**, *10*, 75144–75157. [CrossRef]

25. Guarasci, R.; Minutolo, A.; Damiano, E.; De Pietro, G.; Fujita, H.; Esposito, M. ELECTRA for neural coreference resolution in Italian. *IEEE Access* **2021**, *9*, 115643–115654. [CrossRef]
26. Kuo, C.C.; Chen, K.Y. Toward zero-shot and zero-resource multilingual question answering. *IEEE Access* **2022**, *10*, 99754–99761. [CrossRef]
27. Italian Corpus of Linguistic Acceptability (Repository). Available online: https://paperswithcode.com/dataset/itacola (accessed on 24 April 2024).
28. Bonetti, F.; Leonardelli, E.; Trotta, D.; Raffaele, G.; Tonelli, S. 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, Marseille, France, 20–25 June 2022; pp. 1740–1750.
29. Cho, H. Analyzing ChatGPT's Judgments on Nativelikeness of Sentences Written by English Native Speakers and Korean EFL Learners. *Multimed.-Assist. Lang. Learn.* **2023**, *26*, 9–32.
30. Qiu, Z.; Duan, X.; Cai, Z.G. Grammaticality Representation in ChatGPT as Compared to Linguists and Laypeople. *arXiv* **2024**, arXiv:2406.11116.
31. Ranaldi, L.; Pucci, G. Knowing knowledge: Epistemological study of knowledge in transformers. *Appl. Sci.* **2023**, *13*, 677. [CrossRef]
32. Linzen, T.; Oseki, Y. The reliability of acceptability judgments across languages. *Glossa J. Gen. Linguist.* **2018**, *3*, 100.
33. Cherniavskii, D.; Tulchinskii, E.; Mikhailov, V.; Proskurina, I.; Kushnareva, L.; Artemova, E.; Barannikov, S.; Piontkovskaya, I.; Piontkovski, D.; Burnaev, E. Acceptability judgements via examining the topology of attention maps. *arXiv* **2022**. arXiv:2205.09630.
34. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels, Belgium, 1 November 2018; pp. 353–355.
35. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020.
36. Wang, W.; Bi, B.; Yan, M.; Wu, C.; Xia, J.; Bao, Z.; Peng, L.; Si, L. StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020.
37. Wang, S.; Fang, H.; Khabsa, M.; Mao, H.; Ma, H. Entailment as Few-Shot Learner. *arXiv* **2021**, arXiv:2104.14690.
38. Someya, T.; Oseki, Y. JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs. In Proceedings of the Findings of the Association for Computational Linguistics: EACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 1536–1549.
39. Xiang, B.; Yang, C.; Li, Y.; Warstadt, A.; Kann, K. CLiMP: A Benchmark for Chinese Language Model Evaluation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 2784–2790. [CrossRef]
40. Hu, H.; Zhang, Z.; Huang, W.; Lai, J.Y.K.; Li, A.; Patterson, Y.; Huang, J.; Zhang, P.; Lin, C.J.C.; Wang, R. Revisiting Acceptability Judgements. *arXiv* **2023**, arXiv:2305.14091.
41. Sprouse, J.; Almeida, D. The empirical status of data in syntax: A reply to Gibson and Fedorenko. *Lang. Cogn. Processes* **2013**, *28*, 222–228. [CrossRef]
42. Lau, J.H.; Clark, A.; Lappin, S. Measuring gradience in speakers' grammaticality judgements. In Proceedings of the Annual Meeting of the Cognitive Science Society, Quebec City, QC, Canada, 23–26 July 2014; Volume 36.
43. Marvin, R.; Linzen, T. Targeted Syntactic Evaluation of Language Models. *arXiv* **2019**, arXiv:1808.09031.
44. Feldhausen, I.; Buchczyk, S. Testing the reliability of acceptability judgments for subjunctive obviation in French. In Proceedings of the Going Romance 2020, Online, 25–27 November 2020.
45. Chen, Z.; Xu, Y.; Xie, Z. Assessing introspective linguistic judgments quantitatively: The case of The Syntax of Chinese. *J. East Asian Linguist.* **2020**, *29*, 311–336. [CrossRef]
46. Brunato, D.; Chesi, C.; Dell'Orletta, F.; Montemagni, S.; Venturi, G.; Zamparelli, R. AcCompl-it @ EVALITA2020: Overview of the Acceptability & Complexity Evaluation Task for Italian. In Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online Event, 17 December 2020; Basile, V., Croce, D., Maro, M.D., Passaro, L.C., Eds.; CEUR Workshop Proceedings; Volume 2765.
47. Guarasci, R.; Buonaiuto, G.; De Pietro, G.; Esposito, M. Applying Variational Quantum Classifier on Acceptability Judgements: A QNLP experiment. *Numer. Comput. Theory Algorithms NUMTA* **2023**, 116.
48. Sprouse, J.; Schütze, C.; Almeida, D. Assessing the reliability of journal data in syntax: Linguistic Inquiry 2001–2010. *Lingua* **2013**, *134*, 219–248. [CrossRef]
49. Snow, R.; O'connor, B.; Jurafsky, D.; Ng, A.Y. Cheap and fast–but is it good? Evaluating non-expert annotations for natural language tasks. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; pp. 254–263.
50. Lau, J.H.; Clark, A.; Lappin, S. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cogn. Sci.* **2017**, *41*, 1202–1241. [CrossRef] [PubMed]
51. Fornaciari, T.; Cagnina, L.; Rosso, P.; Poesio, M. Fake opinion detection: How similar are crowdsourced datasets to real data? *Lang. Resour. Eval.* **2020**, *54*, 1019–1058. [CrossRef]

52. Ott, M.; Cardie, C.; Hancock, J.T. Negative deceptive opinion spam. In Proceedings of the 2013 Conference of the north American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 497–501.
53. Guarasci, R.; Catelli, R.; Esposito, M. Classifying deceptive reviews for the cultural heritage domain: A lexicon-based approach for the Italian language. *Expert Syst. Appl.* **2024**, *252*, 124131. [CrossRef]
54. Ruan, N.; Deng, R.; Su, C. GADM: Manual fake review detection for O2O commercial platforms. *Comput. Secur.* **2020**, *88*, 101657. [CrossRef]
55. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
56. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Processing Syst.* **2017**, *30.*
57. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to fine-tune bert for text classification? In Proceedings of the China national conference on Chinese computational linguistics, Kunming, China, 18–20 October 2019; Springer: Cham, Switzerland, 2019; pp. 194–206.
58. dbmdz BERT and ELECTRA Models. Available online: https://huggingface.co/dbmdz/bert-base-italian-xxl-cased (accessed on 20 June 2024).
59. Open Source Project on Multilingual Resources for Machine Learning (OSCAR). Available online: https://traces1.inria.fr/oscar/ (accessed on 20 June 2024).
60. OPUS Corpora Collection. Available online: http://opus.nlpl.eu/ (accessed on 20 June 2024).
61. Rogers, A.; Kovaleva, O.; Rumshisky, A. A primer in bertology: What we know about how bert works. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 842–866. [CrossRef]
62. Electra Base Iitalian XXL Cased. Available online: https://huggingface.co/dbmdz/electra-base-italian-xxl-cased-discriminator (accessed on 20 June 2024).
63. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; pp. 2227–2237.
64. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
65. Liu, H. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua* **2010**, *120*, 1567–1578. [CrossRef]
66. Di Liello, L.; Gabburo, M.; Moschitti, A. Efficient pre-training objectives for transformers. *arXiv* **2021**, arXiv:2104.09694.
67. Margiotta, V. Modeling and Classifying Textual Data through Transformer-Based Architecture: A Comparative Approach in Natural Language Processing. Ph.D. Thesis, Politecnico di Torino, Turin, Italy, 2021.
68. Tepecik, A.; Demir, E. Emotion Detection with Pre-Trained Language Models BERT and ELECTRA Analysis of Turkish Data. *Intell. Methods Eng. Sci.* **2024**, *3*, 7–12.
69. Warstadt, A.; Bowman, S.R. Grammatical Analysis of Pretrained Sentence Encoders with Acceptability Judgments. *arXiv* **2019**, arXiv:1901.03438.
70. Burzio, L. *Italian Syntax: A Government-Binding Approach*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1986; Volume 1.
71. Manning, C.D.; Sag, I.A. Argument structure, valence, and binding. *Nord. J. Linguist.* **1998**, *21*, 107–144. [CrossRef]
72. Chesi, C. An efficient Trie for binding (and movement). *Comput. Linguist. Clic-It* **2018**, 105.
73. Brunato, D.; De Mattei, L.; Dell'Orletta, F.; Iavarone, B.; Venturi, G. Is this Sentence Difficult? Do you Agree? In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2690–2699.
74. Varda, A.G.d.; Marelli, M. Data-driven Cross-lingual Syntax: An Agreement Study with Massively Multilingual Models. *Comput. Linguist.* **2023**, *49*, 261–299. [CrossRef]
75. Marulli, F.; Pota, M.; Esposito, M.; Maisto, A.; Guarasci, R. Tuning syntaxnet for pos tagging italian sentences. *Lect. Notes Data Eng. Commun. Technol.* **2018**, *13*, 314–324. [CrossRef]
76. Warstadt, A.; Parrish, A.; Liu, H.; Mohananey, A.; Peng, W.; Wang, S.F.; Bowman, S.R. BLiMP: The benchmark of linguistic minimal pairs for English. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 377–392. [CrossRef]
77. Buonaiuto, G.; Guarasci, R.; Minutolo, A.; De Pietro, G.; Esposito, M. Quantum transfer learning for acceptability judgements. *Quantum Mach. Intell.* **2024**, *6*, 13. [CrossRef]
78. Li, L.; Li, Z.; Chen, Y.; Li, S.; Zhou, G. Prompt-Free Few-Shot Learning with ELECTRA for Acceptability Judgment. In Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, Foshan, China, 12–15 October 2023; Springer: Cham, Switzerland, 2023; pp. 43–54.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.