Contents lists available at *Dergipark*

Journal of Scientific Reports-A

journal homepage: *https://dergipark.org.tr/tr/pub/jsr-a*

E-ISSN: 2687-6167 ***Number 61, June 2025***

RESEARCH ARTICLE

Receive Date: 11.12.2024 *Accepted Date: 05.05.2025*

Performance analysis of the most downloaded Turkish and English language models on the Hugging-Face platform

İnayet Hakkı Çizmeci[a,*], Kerem Gencer[b]

[a]"*Afyon Kocatepe University, Department of Computer Engineering, Afyonkarahisar 03030, Türkiye,*" *ORCID : 0000-0001-6202-4807*
[b]"*Afyon Kocatepe University, Department of Computer Engineering, Afyonkarahisar 03030, Türkiye,*" *ORCID : 0000-0002-2914-1056*

Abstract

This study analyzes the performance of the most popularly downloaded language models on the Hugging Face platform. For this purpose, the five most downloaded language models in Turkish and English were used. The analysis was evaluated in three phases. These stages were contextual learning, question and answer, and expert evaluation. ARC, Turkish sentiment analysis, Hellaswag, and MMLU datasets were used for contextual learning. For the question-and-answer test, the models trained with the text file created were asked questions from the text. Finally, six experts evaluated the answers given by the models from the developed mobile application. F1 score was used for context evaluation, Rouge-1, Rouge-2, and Rouge-L metrics were used for question and answer, and Elo and TrueSkill metrics were used for expert evaluations. The correlations of these metrics were calculated, and it was seen that there was a correlation of 0.74 between expert evaluations and question-answer performances. It was also observed that learning in context and question-answering performances were not correlated. When the language models were evaluated in general, the timpal0l/mdeberta-v3-base-squad2 language model performed the best. Turkish and English language models performed best on the sentiment analysis dataset with an F1 score above 0.85.

Keywords: Language Models; Fine-tune; Hugging Face; LLM

* Corresponding author. Tel.: +90 272 218 2364
E-mail address: icizmeci@aku.edu.tr

1. Introduction

In today's technology, it has become widespread to produce solutions by developing machine-based models (ML) for solving problems. Unique open-source platforms have been developed for sharing and developing these solutions. The primary reason for using open-source platforms is that machine learning models are large and expensive. To solve this problem, researchers have attempted to adapt pretrained models to open-source platforms. The presentation of model structures, including pre-trained weights, structures, and documentation, has made these platforms popular [1]. The most popular of these platforms is the 'Hugging Face Hub' [2]. The ecosystem of the Hugging Face Hub platform is given in Fig 1.

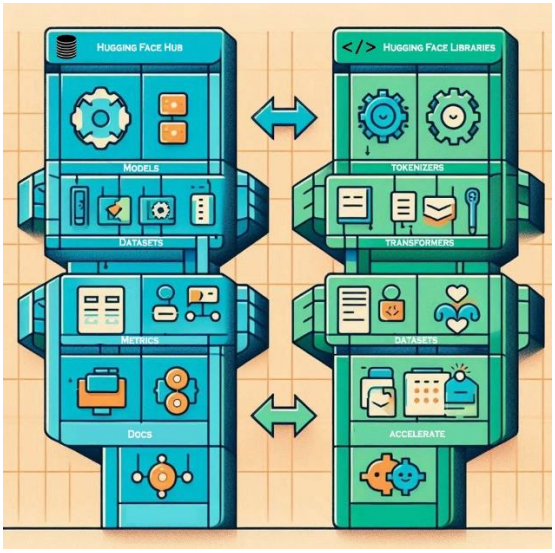

Fig. 1 Main components of the Hugging Face ecosystem [3]

The Hugging Face (HF) platform has become a hub for pretrained models. The HF platform allows language model developers to store the models they create and researchers to develop applications based on these models [4]. The HF platform makes frameworks such as Keras, Tensorflow, and PyTorch available to researchers via API [5]. As of August 2024, it hosts more than 820 thousand models, over 190 thousand data sets, and over 230 thousand demo applications [6]. It offers many public and online arguments that make the platform stand out. It allows everyone to develop natural language processing-based models and applications quickly. Large language models (LLM) such as GPT2, developed by OpenAI; BERT, developed by Google; and LLMA 3, developed by META, are offered to users via HF. If we make a metaphor for the HF platform, teaching mathematics to an illiterate student will take a lot of time and be difficult. However, for a literate student, learning mathematics will be both easier and less time-consuming. HF can be considered a platform through which literate students can obtain information.

Table 1. Some studies on language models

Study Name	Author(s)	Year	Subject	Results
Web Application for Solving Complex Artificial Intelligence Problems	Shen et al. [7]	2023	Use of the ChatGPT model	A web application was developed using ChatGPT, and artificial intelligence problems were solved.

Safety Analysis of Hugging Face Models	Kathikar et al. [8]	2023	Security vulnerabilities of the models	It is stated that the vulnerabilities of the models are 35.98%.
Predicting Early Diagnosis of Mental Disorders	Pourkeyvan et al. [9]	2024	BERT-based fine-tuned models	The prediction was made with four different fine-tuned models of BERT.
Quantitative Analysis of Hugging Face Models	Osborne et al. [10]	2024	Number of downloads of models and usage habits	The number of downloads of 70% of the models was found to be 0, and the number of downloads of 99% was found to be 1.
User and Community Analysis of Models	Castaño et al. [11]	2024	Communities and Model Care Situations	The communities' models, usage frameworks, and maintenance processes were analysed.
Performance Comparison of Turkish Language Models	Dogan et al. [12]	2024	Learning and question-answer performance of Turkish language models in context	In this context, it was found that learning and question-answering capabilities are not significantly related.

Platforms such as LLM Leaderboard [13] and datasets such as BigBench [14] and Big Glue [15] are used to compare the capabilities of language models. However, these platforms and data sets do not include Turkish language models and data sets. To address the deficiencies mentioned in this study, five Turkish and English language models popularly downloaded and fine-tuned on the HF platform were used. Structures such as the artificial intelligence evaluation scale [16] were not used to measure the effectiveness of these models. Although these scales provide an idea about the models, they are not sufficiently evaluated in detail. Therefore, more detailed evaluation and contextual learning approaches are required [17]. Contextual learning, expert evaluation, and question-answer methods were used to evaluate the effectiveness of the models. These evaluations were in the form of contextual learning, expert evaluation, and question-answer. A diagram summarizing the evaluation of the language models for this article is given in Fig. 2.

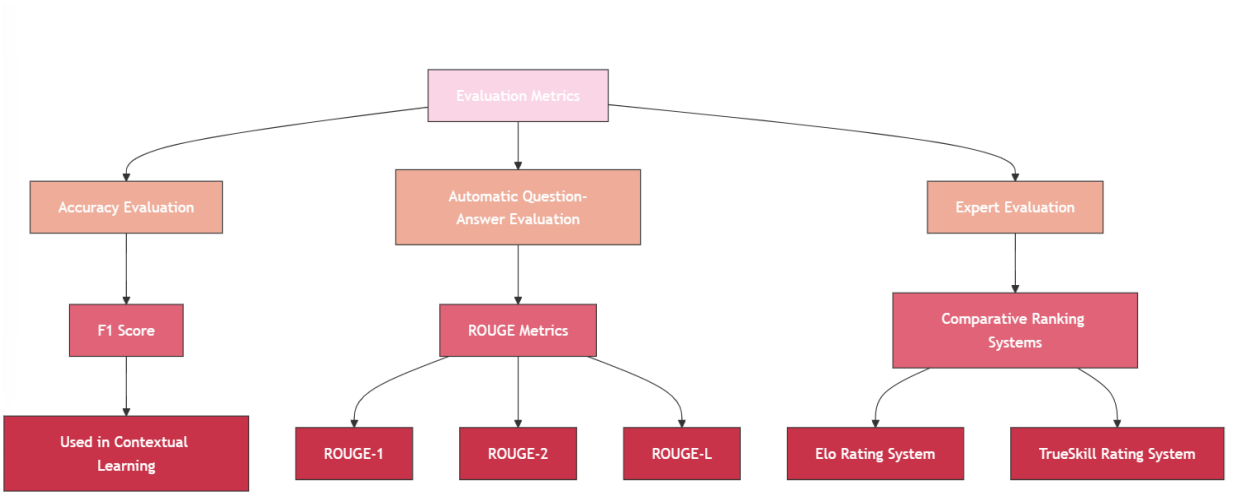

Fig. 2. Diagram summarizing the general structure of the article

The structure of this paper is as follows: In the first section, the purpose of the HF platform and the use of the language models it hosts are explained. The second section gives information about the data sets and methods used to analyze the models. The findings section presents the evaluation of the data obtained in the analysis, and the final section presents a discussion of the results.

2. Materials and Methods

2. 1. Language model (LM)

Language is the most important ability of humans to communicate. Unless the capabilities of machines are improved, they cannot communicate with and understand humans. Experts are constantly researching language models to achieve this goal. This research has led to language modeling methods. The missing parts in language modeling are provided by prediction. Four different methods are used for this modeling [18]. These methods are as follows:

- Statistical Language Model (SLM)
- Neural Language Model (NLM)
- Pre-Trained Language Model (PLM)
- Large Language Model (LLM)

In this study, pre-trained language models were used on the HF platform. The language models used are the platform's top five most downloaded models. As of 12 August 2024, the most downloaded Turkish language models are given in Table 2.

Table 2. Selected Turkish language models

Model Name	Base Model	Description
timpal0l/mdeberta-v3-base-squad2	BERT and RoBERTa	BERT and RoBERTa models were obtained by developing [19].
savasy/bert-base-turkish-squad	BERT	The capacity of the BERT-based model was increased by fine-tuning [20].
incidelen/bert-base-turkish-cased-qa	BERT	boun-tabi/squadtr (Budur et al., 2024) is a model developed using the dataset [21]
yunusemreemik/logo-qna-model	BERT	It is a model inspired by the Savasy/bert-base-turkish-squad model [22].
ozcangundes/mt5-multitask-qa-qg-turkish	Google T5-small	Google's multilingual T5-small model was fine-tuned with the Turkish question-answering dataset [23].

The most downloaded English language models on the same date are given in Table 3.

Table 3. Selected English language models

Model Name	Base Model	Description
deepset/roberta-base-squad2	RoBERTa	It is a model based on the Roberta-base model but gives faster results than this model [24].
bert-large-uncased-whole-word-masking-finetuned-squad	BERT	It is a fine-tuned language model based on the BERT language model developed by Google [25].
distilbert/distilbert-base-cased-distilled-squad	BERT	It is obtained by developing the BERT language model. It stands out with its 60% faster operation and smaller size [26].
distilbert/distilbert-base-uncased-distilled-squad	BERT	It was obtained by developing the BERT language model. 40% fewer parameters were used to make it work faster [26].

phiyodr/bert-large-finetuned-squad2	BERT	Using the large language model, BERT was fine-tuned on SQuAD2.0 [25].

2.2. Data set selection

To analyze the context learning performance of the selected language models, the ARC [27], Turkish Sentiment Analysis [28], Hellaswag, and MMLU [29] datasets were used. For the context evaluation of the models, data were used in both Turkish and English. English datasets were translated to English using the Helsinki-NLP/opus-mt-tc-big-en-tr [30] language model, and Turkish datasets were translated to English using the Helsinki-NLP/opus-mt-tc-big-en [30]. No operations, such as removing meaningless data or structuring, were performed on the existing data. Raw data was used. 200 question-answer data were used for testing in all models. The 200-test data were randomly selected. An example question-answer data set is presented in Table 4.

ARC is a multiple-choice question-answer dataset, easy and hard. The hard section contains difficult questions that require reasoning. Turkish sentiment analysis dataset contains positive, negative, and neutral sentences from various data sources. Hellaswag is a comprehensive dataset that measures the ability of natural language processing systems to complete sentences in a meaningful and logical manner. MMLU is a dataset designed to measure machines' knowledge and reasoning abilities. It contains multiple-choice questions from 57 different fields and topics.

Table 4. Question and answer example

Questions	Choices	Correct Answer
Stars are usually classified according to their brightness as seen in the night sky. Stars can be classified in many other ways. Which of these is least helpful in classifying stars?	A) visible colour, B) composition, C) surface texture D) temperature	C) surface texture
How long does it take for the Earth to rotate on its axis 7 times?	A) one day, B) a week, C) one month, D)one year	B) a week,

A text file containing information about the meta-heuristic algorithms was used for the question-answering performance. The Turkish version of this text file was tested on Turkish language models, and the English version was tested on English language models. A part of this text file is given in Table 5.

Table 5. Part of the Turkish and English texts

Turkish	English
### METASEZGİSEL ALGORİTMALAR	###METAHEURISTIC ALGORITHMS
### Karınca Koloni Algoritması	###Ant Colony Algorithm
Karıncalar besin kaynakları ile evlerinin arasındaki yolları belirlemektedir. İlk olarak geçen karınca feromon adı verilen koku yaymaktadır. Eğer yol kısa ise koku yoğun olmaktadır. Bu durum diğer karıncaların bu yoldan devam etmesini sağlamaktadır. Kesişen yol olursa koku yoğunluğuna göre rastgele seçim olmaktadır.	Ants mark the paths between food sources and their homes. The first ant to pass emits an odor called pheromone. If the path is short, the odor is intense. This ensures that other ants continue this path. If an intersecting path exists, a random selection is made according to the odor intensity.
### Bakteriyel arama besin arama optimizasyonu	###Bacterial search food search optimization
Ekolü bakterilerinin besin arama hareketlerinden esinlenilmiştir. Bakteriler beslenme davranışını örnek almışlardır. Bakteri besini ulaştığında salgı yaymakta ve diğer bakteriler de bu saygıya doğru grup olarak hareket	The foraging behavior of bacteria inspired the school. Bacteria are modeled on feeding behavior. When the bacterium reaches the food, it emits a secretion, and the other bacteria move as a

etmekte bir…	group towards this respect…

Questions were prepared from the text for each model's question-and-answer performance. The answers given automatically to these questions were compared with the answers the experts gave through voting. Examples of the question-and-answer dataset are given in Table 6.

Table 6. Sample questions and answers to the question-and-answer dataset

Turkish Questions	English Questions	Turkish Answers	English Answers
1- Dağ ceylanı optimizasyonunda 4 ana faktör nedir?	1- What are the 4 main factors in mountain gazelle optimization?	Bekar erkekler sürüsü, bölgesel erkekler, annelik sürüleri, yalnızlar	Swarm of single men, territorial men, swarm of mothers, loners
2- Ateş böceği sürüsünde fitness değeri neye göre belirlenmektedir?	2- How is the fitness value determined in a firefly swarm?	Parıldama derecelerine göre fitness değerleri belirlenmektedir.	According to the degree of scintillation
3- Feromon adlı salgı kim tarafından salgılanmaktadır?	3-Who secretes the secretion called pheromone?	İlk karınca tarafından salgılanmaktadır	The first ant to pass
4- Gri Kurt algoritması kaç katmandan oluşur?	4- How many layers does the grey wolf algorithm consist of?	4 katmandan oluşur	4 Layers
5- Yarasa algoritmasında yarasalar uzaklıklarını nasıl belirliyorlar?	5- How do bats determine their distance in the bat algorithm?	Seslere göre	According to sounds

2.3. Methods used in the analysis

2.3.1. Learning method in context

Learning in context is defined as the response of a language model based on the current context without additional training or with very little data. This shows the fast learning and adaptability of model [31]. ARC, Turkish sentiment analysis, Hellaswag, and MMLU datasets were used to determine the contextual capabilities of the popular models selected for this study. F1 scores were calculated based on the accuracy of the models' responses to these datasets. Thus, the strengths and weaknesses of the language models were determined.

2.3.2. Question and answer method

The model's ability to answer automatically is evaluated in the question-and-answer method. For this evaluation, the model was trained with a text file and tested with questions generated from this text. The answers given by the models were analyzed by comparing them with the reference answers. This analysis used ROUGE-1, ROUGE2, and ROUGE-L metrics [32]. They were also analyzed in terms of word order.

F1-Score: The F1 score, a one-dimensional indicator, has an important place in performance evaluation metrics. It is defined as the harmonic means of precision and recall. F1 score takes a value between 0 and 1. While a value of 1 indicates excellent precision and recall values, a value of 0 indicates the worst performance. F1 score is especially prominent in unbalanced data sets [33] [34].

ROUGE: It is a widely used metric in natural language processing. The metric measures numerically how closely an automatically generated summary matches human-generated reference texts. This metric is typically based on word

overlap. A high ROUGE score indicates that the generated text is more similar to the reference text. The ROUGE can be calculated using different methods depending on the level of detail. The most commonly used examples are [35]:

- ROUGE-N: Based on N-gram overlaps (like ROUGE-1, ROUGE-2)
- ROUGE-L: Based on Longest Common Subsequence (LCS) length.
- ROUGE-S: Based on Skip-bigram overlaps.
- ROUGE-SU: Based on Skip-bigram and unigram overlaps.

2.3.3 Expert assessment

In analyzing the performance of the models, the evaluation of the experts by voting method is important. The two models generate the answer to the question randomly selected from the question pool. The expert compares these two answers and chooses one of the four options. A mobile application was developed for this case. The application was developed using the React Native framework and JavaScript. Firebase database was used. The answers to the questions about the models to be compared are given in the mobile application in Fig. 3. It is not stated which answer belongs to the given model. A blind evaluation system has been created. The same questions were asked by all models. The score table for the model was created according to the experts' answers. If the answer of the model is good, the model receives a (+1) point, whereas the other model receives a (-1) point. If the expert chooses the option where both models are good, both models receive a value of (+1), and if both are bad, both models receive a value of (-1).

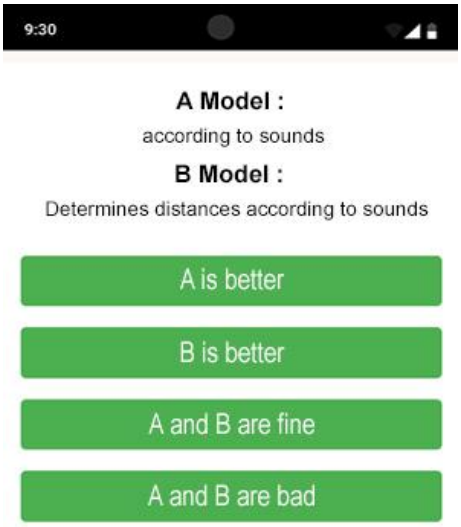

Fig. 3. Mobile App

Six experts can also reveal the semantic differences between the language used by humans and the language produced by the models. In this analysis, Elo [36] and TrueSkill [37] metrics, which are used to evaluate performances in competitive systems, were used. Elo indicates that one model wins, and the other model loses points.

The model with the most points stands out in the evaluation, which starts with a specific score. TrueSkill considers the uncertainties of both sides. Thus, it is considered in cases where A and B are good, or A and B are bad.

2.4 Hardware and Software Used in the Analysis

The experimental study used Google's Collaborative platform for the analysis. The hardware provided by this platform is 40 GB GPU and 107.7 GB storage space. Python was used as the programming language. HF used Transformers and torch libraries to integrate the models into the software. Transformers bring a fast architectural structure for natural language processing, leaving recurrent artificial neural networks behind. This architecture is scaled with training data and model size, enabling more efficient training [38]. PyTorch is an open-source Python-based machine-learning library [39]. The given texts need to be chunked for training. This process is called tokenization. Tokenization is dividing the body given in language models into units. Each unit consists of a token [40]. An example of the code block used for tokenization is given in Fig. 4.

```
model_name = "timpal0l/mdeberta-v3-base-squad2"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForQuestionAnswering.from_pretrained(model_name)
```

Fig. 4. Tokenization process

3. Results

The results of the analyses are given in Table 7. The F1 score was used for accuracy evaluation in contextual learning. ROUGE-1, ROUGE2 and ROUGE-L metrics were used for automatic question-and-answer evaluation, and Elo and TrueSkill were used for expert evaluation. In Elo, each model started with 500 points, while TrueSkill started with 25 points. In the expert evaluation, each expert voted on 70 questions, and 420 votes were used. The correlation matrix of the metrics with each other is shown in Fig. 5. As a result of the analyses, the best result obtained is the Turkish language model A (timpal0l/mdeberta-v3-base-squad2). This language model was developed based on the BERT and RoBERTa language models. Following language model A, language model N (distilbert/distilbert-base-uncased-distilled-squad) performed the second best. Among the other language models, K (deepest/Roberta-base-squad2), L (bert-large-uncased-whole-word-masking-finetuned-squad), and O (phiyodr/bert-large-finetuned-squad2) language models showed the third best performance. The remaining language models could not take the lead in any datasets.

Table 7. Evaluation of models

Criteria	A	B	C	D	E	K	L	M	N	O
ARC (F1 Score)	0.2605	0.2450	0.2559	0.2281	0.2504	0.2778	0.2784	0.2466	**0.2785**	0.2721
Turkish Sentiment Analysis (F1 Score)	**0.9500**	0.9049	0.4251	0.8997	0.8639	0.8486	0.8851	0.8750	0.8845	0.9000
MMLU (F1 Score)	0.2116	0.2183	0.2294	0.2674	0.2305	0.2314	0.2280	0.2289	0.2214	**0.3100**
Hellaswag (F1 Score)	0.1981	0.2170	0.2547	0.2128	0.1967	0.1910	**0.2218**	0.2025	0.1345	0.1799
ROUGE-1	**0.8750**	0.0952	0.3077	0.0952	0.1830	0.7500	0.8000	0.6667	0.7500	0.8000
ROUGE-2	**0.6667**	0.0000	0.1818	0.0000	0.1325	**0.6667**	0.4444	0.5714	**0.6667**	0.4444
ROUGE-L	**0.8750**	0.0952	0.3077	0.0952	0.1830	0.7500	0.8000	0.6667	0.7500	0.8000
ELO	**614**	382	534	407	462	488	514	437	566	593

TrueSkill	**29.5**	20.5	26.5	21.5	23.5	24.5	25.5	22.5	27.5	28.5

A: timpal0l/mdeberta-v3-base-squad2

B: savasy/bert-base-turkish-squad

C: incidelen/bert-base-turkish-cased-qa

D: yunusemreemik/logo-qna-model

E: ozcangundes/mt5-multitask-qa-qg-turkish

K: deepset/roberta-base-squad2

L: bert-large-uncased-whole-word-masking-finetuned-squad

M: distilbert/distilbert-base-cased-distilled-squad

N: distilbert/distilbert-base-uncased-distilled-squad

O: phiyodr/bert-large-finetuned-squad2

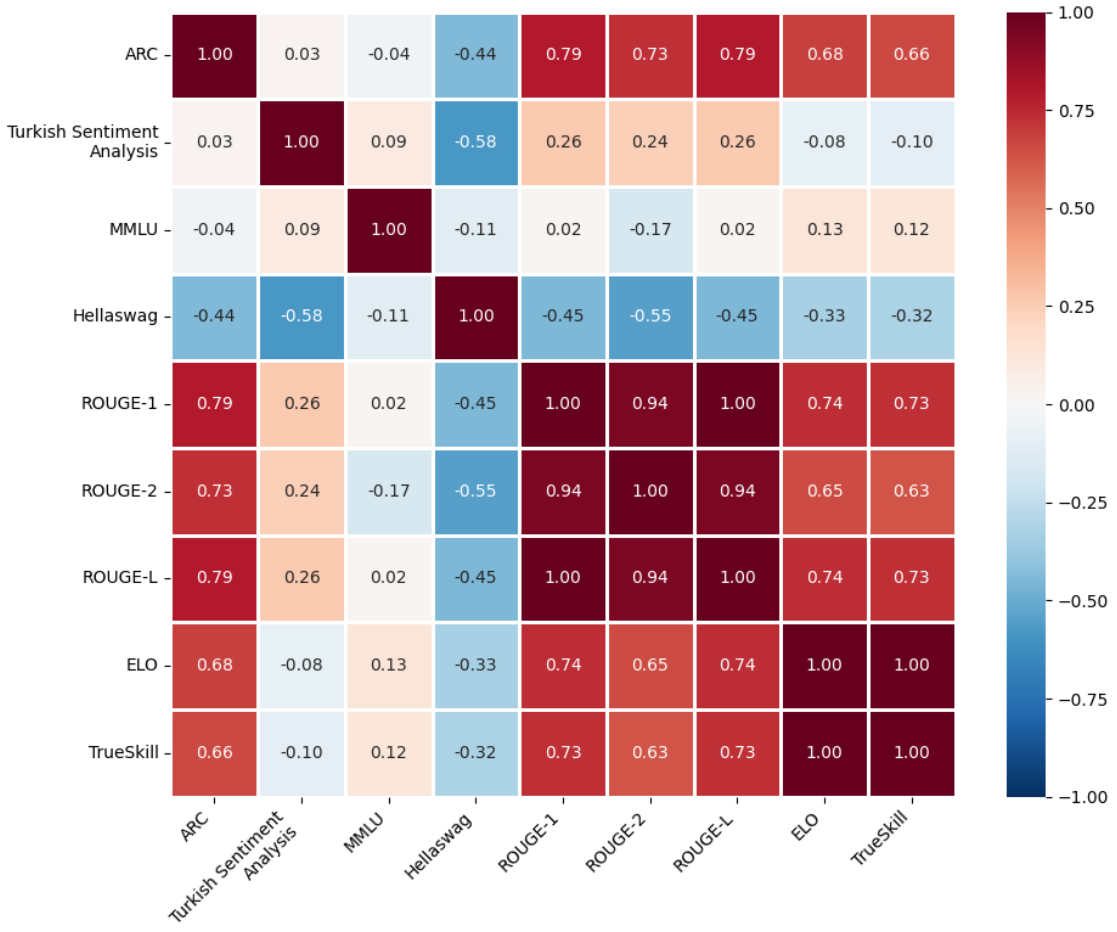

Fig 5. Correlation of Criteria

4. Conclusions and Discussion

The correlation matrix shown in Figure 5 reveals the relationships between different evaluation metrics. According to the analysis results, ROUGE metrics (ROUGE-1, ROUGE-2, and ROUGE-L) have very high correlations among themselves (0.94-1.00). It is also seen that ROUGE metrics show a strong positive correlation (0.73-0.79) with ARC values. An excellent correlation (1.00) is observed between expert evaluation metrics Elo and TrueSkill, indicating that the two methods produce similar results. Elo and TrueSkill also exhibit strong positive correlations (0.63-0.74) with ROUGE metrics. This proves that there is a significant agreement between automatic evaluation and expert evaluation. On the other hand, negative correlations (between -0.11 and -0.58) are observed between Hellaswag and other metrics. These findings suggest that Hellaswag, which measures the ability to learn in context, may be inversely related to question-answering performance. The MMLU metric showed weak correlations with other metrics (between -0.17 and 0.13), indicating that models that perform well on multiple-choice logic and comprehension tests may not perform as well on question-answering tasks. There is a low-to-moderate correlation (0.24-0.26) between Turkish Sentiment Analysis performance and ROUGE metrics, indicating that model performances on sentiment analysis and question-answering tasks may be partially related, but this relationship is not strong.

When we examined the most popular language models on Hugging Face and evaluated the results, we observed something different from what we expected. The most striking outcome was the success of the Turkish model timpal0l/mdeberta-v3-base-squad2. This result demonstrates the importance of combining the strengths of the BERT and RoBERTa models. Moreover, this model provided similarly good results regardless of the measurement method used.

On the other hand, we noticed a big difference between sentiment analysis and question-answer. Although the models achieved F1 scores above 0.85 in sentiment analysis, they scored lower in question-answer. This reminded us that we must select different models for different tasks. Question-answer is a much more complex task than sentiment analysis. In particular, the incidelen/bert-base-turkish-cased-qa model fell behind the others with a low F1 score of 0.42. This showed us that some BERT-based models have limitations in Turkish Question-Answer.

We were surprised to see that different measurement methods such as ROUGE-L, TrueSkill and Elo gave similar results. Therefore, there is less difference between automatic evaluation and expert evaluation than we thought. However, interestingly, we did not find a relationship between the success of a model in context learning and its success in question-answering. This was the case for both the Turkish and English models.

We learned lessons from these results for both researchers and practitioners. For researchers, it became clear that it is important to use different metrics for different tasks rather than a single metric when testing language models. Similar trends in Turkish and English models suggest that some results may be language-independent.

For those working on Turkish natural language projects, we provide practical information on which model to choose for which task. Although mdeberta-v3-base-squad2 is a good choice for general language understanding, other models may be more suitable for tasks focused on sentiment analysis.Finally, the size of the datasets we used in this study was a limitation.

In future, we plan to work with larger and more diverse datasets to increase the reliability of our results. In addition, investigating the reasons for the performance gap between sentiment analysis and question-answering can help us develop better models. Also, the transferability of fine-tuning across languages is an important area of research for multilingual applications.

Acknowledgements

The study did not receive specific financing from any grant agencies in the public, commercial, or non-profit sectors.

References

[1] J. Jones, W. Jiang, N. Synovic, G. Thiruvathukal, and J. Davis, "What do we know about Hugging Face? A systematic literature review and quantitative validation of qualitative claims," in *Proc. of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. 2024, pp. 13–24.
[2] A. Ait, J. L. C. Izquierdo and J. Cabot, "HFCommunity: A Tool to Analyze the Hugging Face Hub Community," in *Proc. 2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. Taipa, Macao, 2023, pp. 728-732, doi:10.1109/SANER56733.2023.00080
[3] Z. Hussain, M. Binz, R. Mata *et al.* "A tutorial on open-source large language models for behavioural science," *Behav Res* 56, pp. 8214–8237, 2024. doi:10.3758/s13428-024-02455-8
[4] S. M. Jain, "Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems," *Apress Media LLC*, pp. 51-53, 2022, doi: 10.1007/978-1-4842-8844-3
[5] F. Pepe, V. Nardone, A. Mastropaolo, G. Bavota, G. Canfora, and M. Di Penta, "How do Hugging Face Models Document Datasets, Bias, and Licenses? An Empirical Study," in *Proc. of the 32nd IEEE/ACM International Conference on Program Comprehension*. 2024, pp. 370–381.
[6] Hugging Face Inc., https://Huggingface.Co/ (accessed August. 13, 2024)
[7] Y. Shen, K. Song, X. Tan, D. Li, W. Lu and Y. Zhuang, "HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. Advances in Neural Information," in *Proc. Systems 36*, New Orleans, USA, 2023, pp. 38154—38180, doi: 10.48550/arXiv.2303.17580.
[8] A. Kathikar, A. Nair, B. Lazarine, A. Sachdeva and S. Samtani, "Assessing the Vulnerabilities of the Open-Source Artificial Intelligence (AI) Landscape: A Large-Scale Analysis of the Hugging Face Platform," in *Proc. 2023 IEEE International Conference on Intelligence and Security Informatics*, Charlotte, NC, USA, 2023, pp.1-6, doi: 10.1109/ISI58743.2023.10297271.
[9] A. Pourkeyvan, R. Safa and A. Sorourkhah, "Harnessing the Power of Hugging Face Transformers for Predicting Mental Health Disorders in Social Networks," IEEE Access, 12, pp. 28025-28035, 2024, doi: 10.1109/ACCESS.2024.3366653
[10] C. Osborne, J. Ding and H. R. Kirk, "The AI community building the future? A quantitative analysis of development activity on Hugging Face Hub," *Journal of Computational Social Science*, vol.7, no.1, pp. 2432-2725, 2024, doi https://doi.org/10.1007/s42001-024-00300-8
[11] J. Castaño, M. F. Silverio, X. Franch and J. Bogner, "Analyzing the Evolution and Maintenance of ML Models on Hugging Face," in *Proc. of the 21st International Conference on Mining Software Repositories*, New York, NY, USA, 2024, pp. 607–618, doi: 10.1145/3643991.3644898
[12] E. Dogan, M. E. Uzun, A. Uz, H. Seyrek, A. Zeer, E. Sevi et al. "Performance Comparison of Turkish Language Models," *arXiv e-prints*, 2024, arXiv:2404.17010.
[13] Open llm leaderboard, a hugging face space by huggingfaceh4, https://huggingface.co/open-llm-leaderboard. (accessed August. 13, 2024)
[14] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch and et al, "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models", *Transactions on Machine Learning Research, 2023*.
[15] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding", *arXiv [cs.CL]*, 20-Apr-2018, arXiv preprint arXiv:1804.07461.
[16] M. Perkins, L. Furze, J. Roe, and J. Macvaugh, "The Artificial Intelligence Assessment Scale (AIAS): a framework for ethical integration of generative AI in educational assessment," *Journal of University Teaching and Learning Practice*, 21(6), 2024. doi: 10.53761/q3azde36
[17] C. Gonsalves, "Contextual assessment design in the age of generative AI," *Journal of Learning Development in Higher Education*, (34), 2025. https://doi.org/10.47408/jldhe.vi34.1307
[18] W. X. Zhao, K. Zhao, J. Li, T. Tang, X. Wang and *et al.*, "A survey of large language models", *arXiv [cs.CL]*, 31-Mar-2023, doi: 10.48550/arXiv.2303.18223
[19] P. He and J. Gao, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing', in *Proc. The Eleventh International Conference on Learning Representations*, Kigali, Ruanda, 2023, pp. 1–16.
[20] S. Yildirim, "Fine-tuning Transformer-based encoder for Turkish Language understanding tasks", *arXiv [cs.CL]*, 30-Jan-2024, doi: 10.48550/arXiv.2401.17396
[21] M. İncidelen, Hugging Face Inc., https://Huggingface.Co/Incidelen/Bert-Base-Turkish-Cased-Qa (accessed August 13, 2024)
[22] Y. E. Emik, Hugging Face Inc., https://Huggingface.Co/Yunusemreemik/Logo-Qna-Model (accessed August 13, 2024)
[23] Ö. Gündeş, Hugging Face Inc., https://Huggingface.Co/Ozcangundes/Mt5-Multitask-Qa-Qg-Turkish (accessed August 13, 2024)
[24] Deepset, Hugging Face Inc., https://huggingface.co/deepset/roberta-base-squad2 (accessed August 13, 2024)
[25] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in *Proc. of NAACL-HLT 2019*, Stroudsburg, PA, USA, 2019, pp. 4171–4186.
[26] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", *arXiv [cs.CL]*, 2019, doi:10.48550/arXiv.1910.01108
[27] naytin, Hugging Face Inc. https://huggingface.co/datasets/naytin/ai2_arc_tr (accessed September 12, 2024)
[28] Hugging Face Inc. https://huggingface.co/datasets/winvoker/turkish-sentiment-analysis-dataset (accessed September 12, 2024)
[29] naytin, Hugging Face Inc. https://huggingface.co/datasets/naytin/hellaswag_tr (accessed September 12, 2024)
[30] J. Tiedemann and S. Thottingal, "OPUS-MT Building open translation services for the World", in *Proc. the 22nd Annual Conference of the European Association for Machine Translation*, 2020, pp. 479–480.
[31] Y. Gu, L. Dong, F. Wei, and M. Huang, "Pre-training to learn in context", *arXiv [cs.CL]*, 15-May-2023, doi: 10.48550/arXiv.2305.09137
[32] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries", in *Text summarization branches out*, 2004, pp. 74–81.
[33] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics* 21, 6, 2020. https://doi.org/10.1186/s12864-019-6413-7.

[34] H. Huang, H. Xu, X. Wang and W. Silamu, "Maximum F1-Score Discriminative Training Criterion for Automatic Mispronunciation Detection," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 787-797, April 2015, doi: 10.1109/TASLP.2015.2409733

[35] M. Barbella and G. Tortora, "Rouge metric evaluation for text summarization techniques," *Available at SSRN 4120317. http://dx.doi.org/10.2139/ssrn.4120317*

[36] A. E. Elo, *The Rating of Chessplayers, Past and Present*. New York: Arco Publishing, 1978.

[37] R. Herbrich and T. Graepel, *TrueSkillTM: A Bayesian skill rating system*. Microsoft Research, 2006.

[38] T. Wolf, L. Debut, V. Chaumond, *C.* Delangue, A.Moi, P. Cistac end *et al.*, "Transformers: State-of-the-Art Natural Language Processing*", The 2020 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, USA, pp. 38–45.

[39] Y. Zhang *et al*., "DIALOGPT: Large-scale generative pre-training for conversational response generation", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Online, 2020, doi:10.18653/v1/2020.acl-demos.30

[40] S. Choo and W. Kim, "A study on the evaluation of tokenizer performance in natural language processing", *Appl. Artif. Intell.*, vol. 37, no. 1, Dec. 2023.