Context-Aware Extraction of Quranic References A Hybrid Language Model- and Rule-Based Approach

Alireza Sahebi
Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
alireza.sahebi@sharif.edu

Mohammadmahdi Hemmatyar
Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
mahdi.hemmat@sharif.edu

Ehsaneddin Asgari*
Qatar Computing Research Institute
Doha, Qatar
easgari@hbku.edu.qa

Abstract

With the increasing use of Quranic expressions in online discourse, religious content, and modern Arabic writing, there is a growing need for tools that can automatically and accurately detect references to the Holy Quran. Furthermore, large language models (LLMs) often generate hallucinated or inaccurate Quranic content, highlighting the importance of tools capable of verifying and correcting such outputs. To address these challenges, this paper presents a multi-layered tool for extracting Quranic expressions from arbitrary input text. A central challenge in this task lies in distinguishing between intentional references and incidental lexical overlap with Quranic text. The proposed tool combines an Arabic language model with rule-based techniques to achieve high precision and contextual understanding. The language model identifies expressions likely intended as Quranic references, effectively filtering out irrelevant matches. These candidate expressions are then verified using regular expression patterns to ensure textual accuracy, returning their span in the input text along with the corresponding Surah and verse number. This hybrid framework enables context-sensitive and semantically accurate extraction of Quranic references, supporting applications in digital humanities, Islamic scholarship, and the enhancement of Quranic content presentation in AI-generated text. The tool and corresponding model are accessible at `https://github.com/llm-lab-org/ayah-detect`.

1 Introduction

The Holy Quran is frequently referenced in modern Arabic discourse, creating a need for robust tools that can automatically and accurately detect its verses within arbitrary text. Such a tool can address several critical needs, e.g., enabling large-scale analysis by extracting verse commentaries from classical texts like the Hadith [4] or facilitating the analysis of Quran usage on social media [1]. Such tools can serve to validate content from Large Language Models (LLMs), which are prone to hallucination [9], by clearly tagging authentic Quranic verses and separating them from non-Quranic text.

A key difficulty in this task is distinguishing intentional Quranic quotations from incidental lexical overlap (similar to intent recognition in text reuse [6]). A naive string-matching approach yields too

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Muslims in ML.

many false positives, failing to capture the user's intent [5]. To address this, we propose a new hybrid framework that takes advantage of language modeling and rule-based techniques. In this framework, a fine-tuned Arabic language model first identifies candidate expressions that are semantically likely to be Quranic references, filtering out irrelevant matches. These candidates are then validated for textual accuracy using precise regular expression (regex) patterns [2], which return the exact match, its span in the input text, and its corresponding Surah and verse number.

This paper presents a method for detecting Quranic expressions in free-form Arabic text, with the following key contributions: **(i) Integrated approach:** The method combines an Arabic language model for context-aware identification with a rule-based regex matcher to enhance precision. **(ii) Reference disambiguation:** It addresses the challenge of distinguishing intentional Quranic quotations from incidental lexical overlap, improving semantic accuracy. **(iii) Publication of the implementation:** The resulting tool will be publicly available to support applications in Quranic verse tagging, digital humanities, and the evaluation of LLM-generated content.

2 Related Works

Early and notable efforts have focused on developing intelligent matching algorithms that go beyond naive string comparison. For instance, Shahmohammadi et al. [13] proposed a two-stage method that first uses a pattern-matching algorithm to identify potential Quranic words, and then applies a filter on their numerical representations to confirm if they form a genuine verse by detecting specific indexing sequences. The tool developed for this framework was not released publicly. The algorithm by Abokhodair et al. [1] splits tweets into sentences and compares them against a list of Quranic verses to find full or partial matches. To reduce false positives from common phrases, it only considers matches that are at least three words long. Similar to previous work, this tool has not been made available to the public. The QDetect system, introduced by El-Beltagy and Rafea [5], uses a two-step algorithm. First, it pre-processes the entire Quran into an efficient, tree-like data structure made of linked hash tables. Then, to find verses in a new text, it scans the text word-by-word, traversing the pre-built structure to find the longest possible match in an efficient manner. This tool represents a significant step forward by providing a publicly available system that has the ability for error correction of misspelled, missing, or wrong words. Recent work by Alam et al. [3] used BERT models to classify Quranic verses as authentic or not, achieving high accuracy by training on a dataset with subtly altered verses. However, their approach focuses only on detection and does not extract the specific address (Surah and verse number) of the verse in Quran, which is a key function in our system.

In recent years, various Arabic LLMs have been introduced, such as Jais [12], Fanar [15], and SILMA [14]. While these models have advanced performance on downstream tasks like question answering, employing them to extract precise character spans for Quranic expressions and Surah and verse numbers may pose challenges. Existing evaluations of LLMs on information extraction tasks substantiate this difficulty. Han et al. [7] highlight that generative models, such as ChatGPT, struggle with strict span matching when compared to state-of-the-art supervised methods. This suggests that the inherent generative nature of LLMs makes them less effective for tasks requiring rigid substring exactness, such as identifying specific character indices of Quranic expressions in texts.

3 Data

For fine-tuning our language model, we utilized a collection of Hadiths provided by the Computer Research Center of Islamic Sciences (Noor)[2]. This collection is valuable as it contains Hadiths annotated with special tags indicating the presence of Quranic verses. We processed the corpus by normalizing the text and splitting it into words, creating a sequence labeling dataset called Noor dataset. This dataset is designed specifically for fine-tuning a language model to detect Quranic expressions on a token-by-token basis. Each word is tagged with one of three labels: 'B-QUR' (beginning of a Quranic expression), 'I-QUR' (inside a Quranic expression), or 'O' (outside of any Quranic expression). The dataset comprises 70,862 Hadiths, containing 221,931 Quranic expressions in total. Table 1 shows a sample from the dataset. To fine-tune the language model, we partitioned the Hadiths into 80% for training, and 10% each for validation and test sets.

[2]`https://hadith.inoor.ir/ar/home`

Table 1: A dataset sample displaying a part of a Hadith alongside its respective B-I-O annotations.

درجات	تعالي	الله	علي	التوكل	فقال	حسبه	فهو	الله	علي	يتوكل	من	و	تعالي	الله	قول	عن	سالته
O	O	O	O	O	O	I-QUR	I-QUR	I-QUR	I-QUR	I-QUR	I-QUR	B-QUR	O	O	O	O	O

The ground truth for all Quranic references is derived from a standardized version of the Holy Quran[3]. This text serves as the basis for generating the regex patterns used in our rule-based method. Additionally, we employed a list of Arabic stop words to filter out and discard trivial matches composed entirely of high-frequency, low-information words like prepositions and conjunctions.

4 Materials and Methods

Our system is built on a hybrid architecture that combines a sensitive rule-based method with a context-aware language model. These components can be used independently or, for optimal performance, in an integrated workflow.

4.1 Rule-Based Regex Method

Our "Regex method" is a pattern-matching engine that operates in stages. First, it performs rigorous "Text Normalization" on both the input and Quranic text using an extension of CAMeL Tools [11]. This involves removing diacritics, tatweel, some punctuations, and numbers; standardizing characters (e.g., converting variants like ک, ڪ, and ګ to a plain Kaf ك); unifying word orthographies (e.g., داوود and داود); and collapsing whitespace.

Next, we perform "Pattern Generation" from the normalized Quran. This includes creating "Global Patterns" for morphological variants (like ة, ت, and ه and suffixes for third-person masculine plurals وا and و), "Bigram Indexing" to map all two-word sequences to their locations (handle cases where the conjunction و is attached to the next word), and dynamic "Verse-Level Patterns" that expand from bigrams to capture longer phrases in a verse. These patterns are pre-compiled for faster matching.

Finally, the "Search Algorithm" finds bigram matches in the text, uses verse-specific patterns to identify the longest possible expression, and filters out any substring results. Each match is returned with its text span and Quranic address, with a supplementary filter prioritizing results that span consecutive verses.

4.2 Language Model Method

To understand the contextual likelihood of a passage being a Quranic quote, we used a pre-trained Arabic language model.

Fine-Tuning: We used the CAMeLBERT model (specifically, 'bert-base-arabic-camelbert-ca'), a powerful BERT-based model for Arabic [8], to fine-tune on the sequence labeling task using the Hugging Face Transformers library [16] and the Noor dataset we made in the BIO format (Section 3). The training was conducted using the Google Colab for 18 epochs with a learning rate of 2e-5, using AdamW optimizer [10], a linear learning rate scheduler, a weight decay of 0.01, and a batch size of 16.

Inference and Extraction: To perform extraction, the input text first undergoes normalization and tokenization. We handle texts longer than the 512-token context window of CAMeLBERT by using a sliding window with a 150-token overlap. Following this preprocessing, the model predicts a classification label (B-QUR, I-QUR, or O) for every token in the text.

To improve recall, we use a "logit threshold". This mechanism reassigns a token from the Outside label to the next-highest label (Begin or Inside) if the difference in their logit scores is below this threshold. This allows the model to capture potential verses it is less certain about.

[3]Downloaded from `ai.inoor.ir`

Table 2: Performance comparison of our Regex-only and hybrid methods against the QDetect tool [5]. Our hybrid model, denoted as 'LM+Regex_LTX', is evaluated with various logit thresholds (X). The 'Addr' column specifies whether verse address verification is required, while the 'Overlap %' column defines the minimum textual overlap needed for a true positive.

Model	Addr	Overlap %	Precision	Recall	F1-score
QDetect	✓	0	79.90	50.90	62.18
QDetect	✗	0	83.45	52.69	64.60
QDetect	✗	50	46.04	31.48	37.40
LM+Regex_LT1	✓	0	84.84	98.98	91.37
LM+Regex_LT3	✓	0	82.74	99.02	90.15
LM+Regex_LT5	✓	0	79.69	99.03	88.32
LM+Regex_LT0	✓	0	**85.61**	98.95	**91.80**
LM+Regex_LT0	✗	0	**96.96**	99.34	**98.13**
LM+Regex_LT0	✗	50	**80.37**	98.54	**88.53**
Regex-only	✓	0	13.06	**99.05**	23.08
Regex-only	✗	0	14.87	**99.43**	25.87
Regex-only	✗	50	12.26	**98.66**	21.82

4.3 Integrated Hybrid Method

The most effective application of our tool combines the two methods into a streamlined workflow, leveraging the strengths of both. The process begins with "Candidate Selection", where the fine-tuned CAMeLBERT model processes the input text to identify and extract candidate Quranic expressions, providing a contextually aware first pass. Each of these candidates then undergoes a "Validation and Extraction" stage. For this, a slightly larger span of text is selected, including a few surrounding tokens on each side to ensure the full verse is captured. Afterwards, the sensitive rule-based Regex method is run on this smaller, targeted span. This step confirms the candidate Quranic expression, identifies its addresses in the Quran, and provides its corresponding span in the input text.

5 Results

Our tool's performance was evaluated on a held-out test set from the Noor dataset (Section 3). The following details our evaluation protocol and provides a comparative analysis of our model's performance against another tool.

5.1 Evaluation Methodology

For each Hadith in the test set, we compared the expressions extracted by our tool against the ground truth tags. An extracted expression was counted as a "True Positive" (TP) if it had a textual overlap with a ground truth tag's span *and* its identified Quranic address matched the ground truth. We allowed a small tolerance in the verse number (± 3) to account for cases where a single tag spanned multiple consecutive verses in a Surah. A ground truth tag was counted as "False Negative" (FN) if no corresponding TP was found by our tool. An extracted expression was counted as "False Positive" (FP) if it did not correspond to any ground truth tag in the test set. It is important to note that due to the incomplete nature of the dataset's annotations, some of these FPs may in fact be correct detections of untagged verses. From these counts, we calculated standard Precision, Recall, and F1-Score metrics.

5.2 Comparative Performance Analysis

We evaluated our Regex-only and hybrid methods against the QDetect tool [5] on two tasks: "Detection" of the correct text span and "Classification" of the verse address. For the hybrid system, we varied the language model's logit threshold to adjust its sensitivity and impact on recall, with a threshold of 0 being the most conservative.

The results, shown in Table 2, reveal a clear precision-recall trade-off. The table's "Addr" column (✓/✗) toggles whether a correct verse address is required for a True Positive. When this check is off, the "Overlap %" column specifies the minimum text overlap needed to stay confident about the matched extracted expression. While the Regex-only method achieves the highest recall at the cost

Table 3: Examples of False Negative and False Positive cases produced by the LM+Regex_LT0 model on the held-out test set of the Noor dataset. Yellow text indicates Quranic expressions in the Noor dataset that our model failed to detect; green indicates expressions correctly identified by both the Noor dataset and our tool; and blue marks expressions detected by our tool but not annotated as Quranic in the Noor dataset.

Model prediction	example
False Negative	… فَلِذَلِكَ جاءَ الْقَوْلُ: «وَ عِنْدَ جُهَيْنَةَ الْخَبَرُ الْيَقِينُ» فَلِذَلِكَ قَوْلُهُ: «وَ لَوْ تَرىٰ إِذْ فَزِعُوا» اِلَي اخِرِهِ اَوْرَدَهُ الثَّعْلَبِيُّ فِي تَفْسِيرِهِ … سالت الرضا ـ ع ـ عن بِسْمِ اللّهِ ، قال: معني قول القائل بِسْمِ اللّهِ ، اي: اسم علي نفسي بسمة من سمات اللّه ـ عز و جل ـ …
False Positive	… قالَ اَللّهُ عَزَّ وَ جَلَّ فِيهِمْ لاٰ يَعْصُونَ اَللّهَ مٰا أَمَرَهُمْ وَ يَفْعَلُونَ مٰا يُؤْمَرُونَ وَ قالَ اَللّهُ عَزَّ وَ جَلَّ وَ لَهُ مَنْ فِي اَلسَّمٰاوٰاتِ وَ اَلْأَرْضِ وَ مَنْ عِنْدَهُ … … فَقالَتِ اِمْرَأَةُ فِرْعَوْنَ: اُطْلُبُوا لاِبْنِي ظِئْراً وَ لاَ تُحَقِّرُوا أَحَداً فَجَعَلَ لاَ يَقْبَلُ مِنِ اِمْرَأَةٍ مِنْهُنَّ فَقالَتْ أُمُّ مُوسَى لِأُخْتِهِ: قُصِّيهِ اُنْظُرِي أَ تَرَيْنَ لَهُ أَثَراً …

of very low precision, our hybrid system with a logit threshold of 0 (LM+Regex_LT0) delivers the best overall performance. This optimal configuration achieves the highest F1-score, significantly outperforming both the Regex-only method and the QDetect system.

To further validate our approach, we benchmarked our top-performing model, LM+Regex_LT0, against Fanar [15], an Arabic LLM known for its superior performance across various tasks. This comparison utilized a subset of the Noor test dataset comprising 1.6K Hadiths under both zero-shot and few-shot settings. As detailed in Table 4 in the appendix, our model demonstrates a significant performance advantage over Fanar across all evaluation metrics.

5.3 Analysis of Failure Cases

To gain deeper insight into the limitations of the model, we examined the False Negative and False Positive predictions produced by our best-performing configuration (LM+Regex_LT0). Table 3 presents representative examples of each category.

Regarding False Negative errors, the analysis suggests the model can be sensitive to punctuation marks, particularly guillemets. In the first False Negative example in Table 3, an embedded Quranic expression went undetected; however, removing a pair of guillemets allowed the model to correctly identify it. This may be caused by the low frequency of training samples with multiple guillemet-enclosed expressions in close succession. Another potential issue arises when identical Quranic expressions appear in close proximity. In a specific test case, the model successfully identified the first instance of an expression but missed the second identical one. Adjusting the logit threshold to 1 allowed both to be detected, suggesting that the model might benefit from further fine-tuning on samples with repeated verse structures.

In terms of False Positives, some cases appear to stem from incomplete annotations within the dataset. However, some may be caused by textual similarities, like the blue colored expression in the first False Positive example in Table 3. Additionally, False Positives occur when the input text contains paraphrased versions of Quranic verses. In such cases, the language model perceives the semantic similarity, and the subsequent regex matching labels segments which exactly match with Quran text.

6 Discussion and Conclusion

Our results confirm the effectiveness of our hybrid architecture, whose main strength lies in the synergy between its components. The language model acts as a contextual pre-filter, significantly boosting the precision of the Regex method by eliminating false positives from incidental word overlap. While the language model alone is insufficient because it cannot provide the essential Quranic addresses (Surah and verse), our Regex method supplies them, ensuring practical usability. Furthermore, a tunable logit threshold allows recall to be increased for applications requiring more comprehensive detection.

In conclusion, we introduced a novel, hybrid framework that effectively detects Quranic verses by using a language model for contextual filtering and a rule-based engine for validation and addressing. Future work will focus on enhancing the training data with more comprehensive tagging and leveraging more powerful language models. The tool will be made publicly available to support research and development in Quranic NLP.

7 Acknowledgement

We gratefully acknowledge the publicly accessible Hadith collection provided by the Computer Research Center of Islamic Sciences (Noor) at https://hadith.inoor.ir/ar/home, which served as the basis for our language-model fine-tuning. All data were obtained from publicly available pages on the website, and no special access or privileged permissions were required.

References

[1] Norah Abokhodair, AbdelRahim Elmadany, and Walid Magdy. Holy tweets: Exploring the sharing of the quran on twitter. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), October 2020. doi: 10.1145/3415230. URL https://doi.org/10.1145/3415230.

[2] Alfred V Aho. Algorithms for finding patterns in strings, handbook of theoretical computer science (vol. a): algorithms and complexity, 1991. URL https://www.sciencedirect.com/science/article/abs/pii/B9780444880710500102.

[3] Khubaib Amjad Alam, Maryam Khalid, Syed Ahmed Ali, Haroon Mahmood, Qaisar Shafi, Muhammad Haroon, and Zulqarnain Haider. Automated authentication of Quranic verses using BERT (bidirectional encoder representations from transformers) based language models. In Sane Yagi, Sane Yagi, Majdi Sawalha, Bayan Abu Shawar, Abdallah T. AlShdaifat, Norhan Abbas, and Organizers, editors, *Proceedings of the New Horizons in Computational Linguistics for Religious Texts*, pages 59–66, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.clrel-1.6/.

[4] Shatha Altammami and Eric Atwell. Challenging the transformer-based models with a classical Arabic dataset: Quran and Hadith. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1462–1471, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.157/.

[5] Samhaa R. El-Beltagy and Ahmed Rafea. Qdetect: An intelligent tool for detecting quranic verses in any text. *Procedia Computer Science*, 189:374–381, 2021. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2021.05.107. URL https://www.sciencedirect.com/science/article/pii/S1877050921012321. AI in Computational Linguistics.

[6] Greta Franzini, Elisa Franzini, and Michael Büchler. Historical text reuse: What is it?, 2016. URL http://www.etrap.eu/historical-text-re-use/. Accessed: 2025-07-07.

[7] Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*, page 48, 2023.

[8] Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*, 2021. URL https://aclanthology.org/2021.wanlp-1.10v2.pdf.

[9] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL http://dx.doi.org/10.1145/3571730.

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[11] Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. CAMeL tools: An open source python toolkit for Arabic natural language processing. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk,

and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.868/.

[12] Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models, 2023. URL https://arxiv.org/abs/2308.16149.

[13] Mohsen Shahmohammadi, Toktam Alizadeh, Mohammad Habibzadeh Bijani, and Behrouz Minaei. A framework for detecting holy quran inside arabic and persian texts. In *Workshop Organizers*, page 71, 2015. URL https://www.researchgate.net/profile/Thabit-Sabbah/publication/270884459_A_Framework_for_Quranic_Verses_Authenticity_Detection_in_Online_Forum/links/54f753750cf28d6dec9e79fa/A-Framework-for-Quranic-Verses-Authenticity-Detection-in-Online-Forum.pdf.

[14] silma-ai. Silma 9b instruct v1.0. https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0, 2024.

[15] Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. Fanar: An arabic-centric multimodal generative ai platform, 2025. URL https://arxiv.org/abs/2501.13944.

[16] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL https://arxiv.org/abs/1910.03771.

A Evaluation of Fanar LLM

Table 4: Performance comparison of our best model, LM+Regex_LT0, againts the Fanar LLM on 1.6k Hadiths of Noor test dataset

Model	Addr	Overlap %	Precision	Recall	F1-score
Fanar_zero-shot	✓	0	12.21	15.28	13.58
Fanar_zero-shot	✗	0	28.37	38.8	32.78
Fanar_zero-shot	✗	50	1.42	1.78	1.58
Fanar_few-shot	✓	0	11.67	12.27	11.96
Fanar_few-shot	✗	0	35.94	43.28	39.27
Fanar_few-shot	✗	50	10.12	11.61	10.81
LM+Regex_LT0	✓	0	**87.13**	**99.19**	**92.77**
LM+Regex_LT0	✗	0	**97.44**	**99.42**	**98.42**
LM+Regex_LT0	✗	50	**81.37**	**98.62**	**89.17**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the three key contributions of the paper are mentioned in the section 3.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The incompleteness of the dataset's annotations are mentioned in section 5.1.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The procedure to fine-tune the model plus the necessary hyperparameters are mentioned in section 4.2. All the steps necessary to create the regex patterns are mentioned in section 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and trained model will be publicly available after the acceptance of the paper. The Hadith data used for this research can be accessed from the Computer Research Center of Islamic Sciences (Noor) website, as cited in the footnote in section 3.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have mentioned the percentages for data split in section 3 and the hyperparameters in section 4.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The environment which the training is conducted is mentioned in section 4.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The positive impacts are mentioned in abstract and conclusion sections.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer:[Yes]

    Justification: The owners of the data and the pretrained language model are identified in Sections 3 and 4.2, respectively

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The base model is freely available in huggingface, which is mentioned in section 4.2.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.