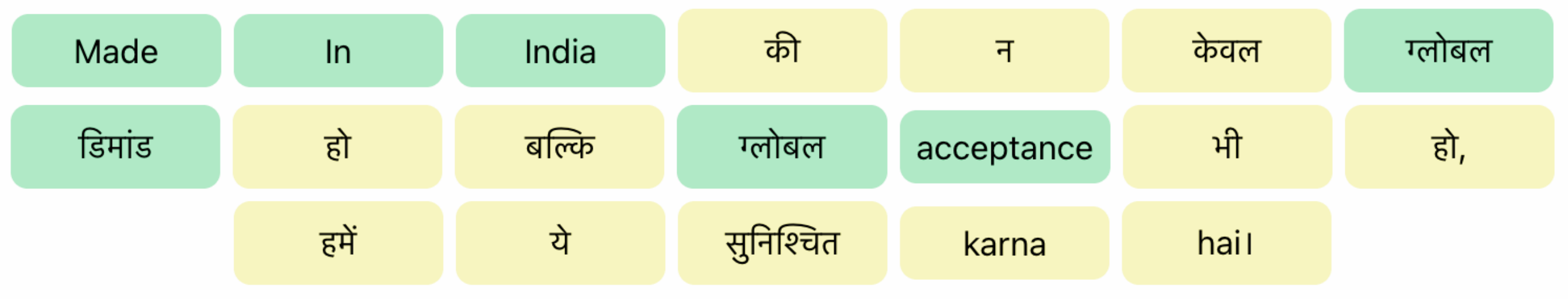
Commentator ✍: A Code-mixed Multilingual Text Annotation Framework

Rajvee Sheth† , Shubh Nisar*, Heenaben Prajapati † , Himanshu Beniwal † , Mayank Singh†

Discipline of Computer Science and Engineering, Indian Institute of Technology Gandhinagar †
Department of Computer Science, North Carolina State University *
{rajvee.sheth, heenaben.prajapati, himanshubeniwal, singh.mayank}@iitgn.ac.in, shubhnisar123@gmail.com

Introduction

Code-Mixed text, two or more languages alternate within a sentence or conversation, is increasingly prevalent in social media and informal communication.


Made In India की न केवल ग्लोबल
डिमांड हो बल्कि ग्लोबल acceptance भी हो,
हमें ये सुनिश्चित karna hai।

Motivation:
Code-mixed language is widely used across social media platforms.
There is a significant shortage of annotated resources for code-mixed languages.
Annotated data is vital for training effective multilingual models and chatbots.

Contributions:

1. Introduced **COMMENTATOR**, a robust framework designed for efficient annotation of code-mixed multilingual text.
2. Evaluated it through a detailed analysis against 5-6 SOTA annotation frameworks.
3. With improved collaboration and efficiency, it reduces annotation time by **5x** for **LID[1] task** and **2x** for **POS[2] task** over the best baseline.

COMMENTATOR

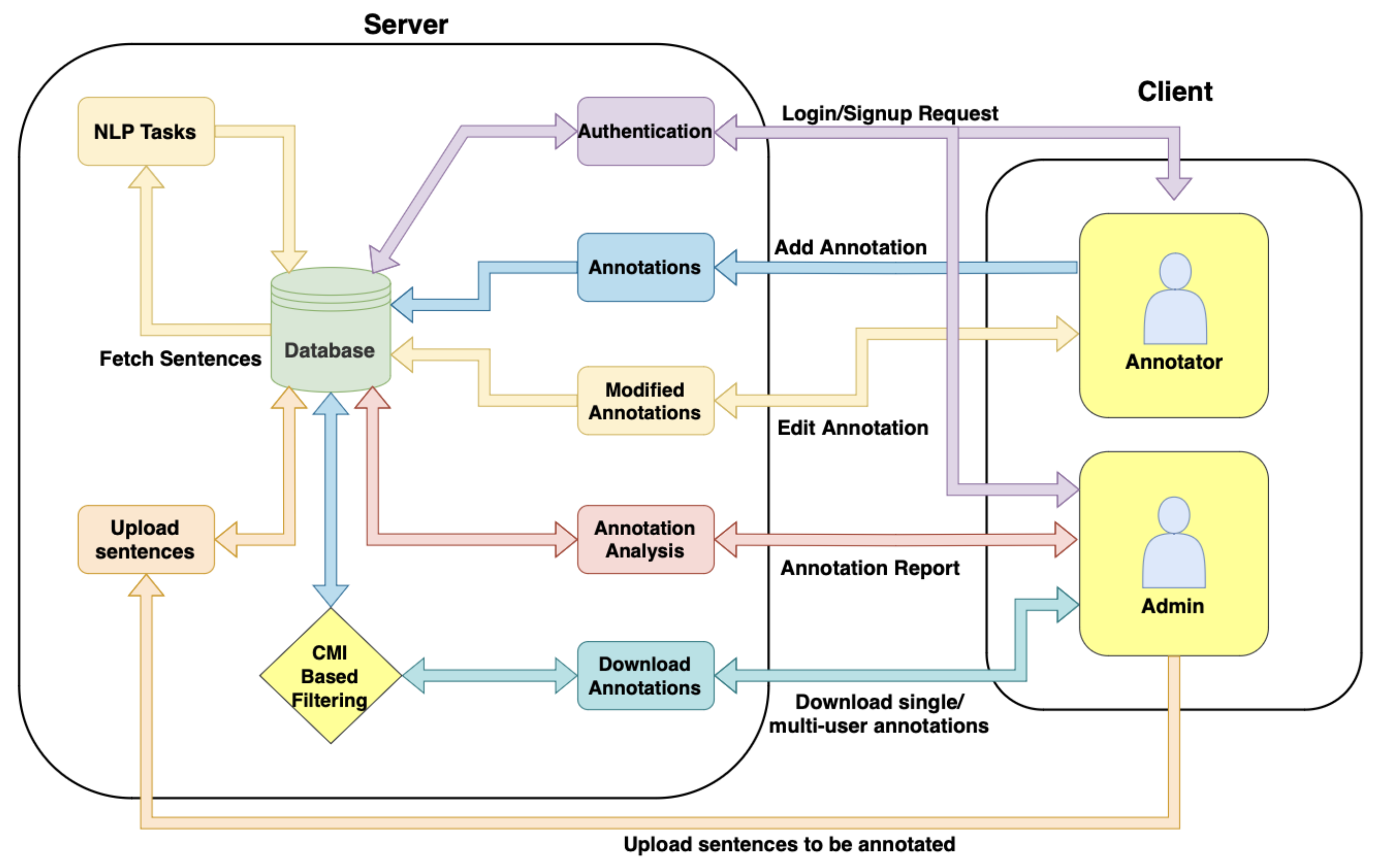
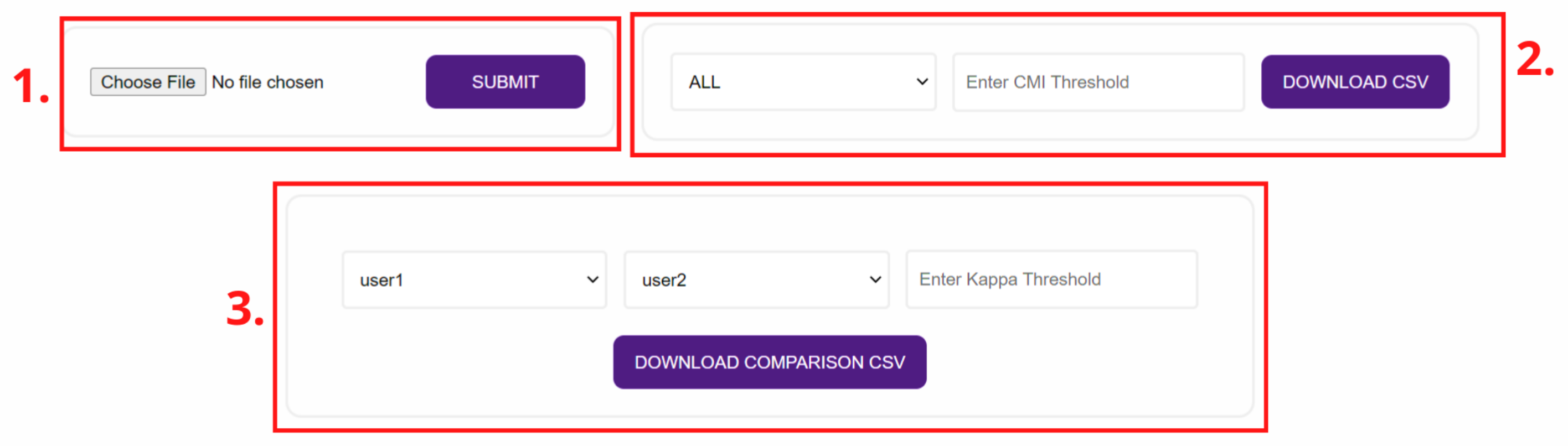


Figure 1: The proposed architecture of Commentator.

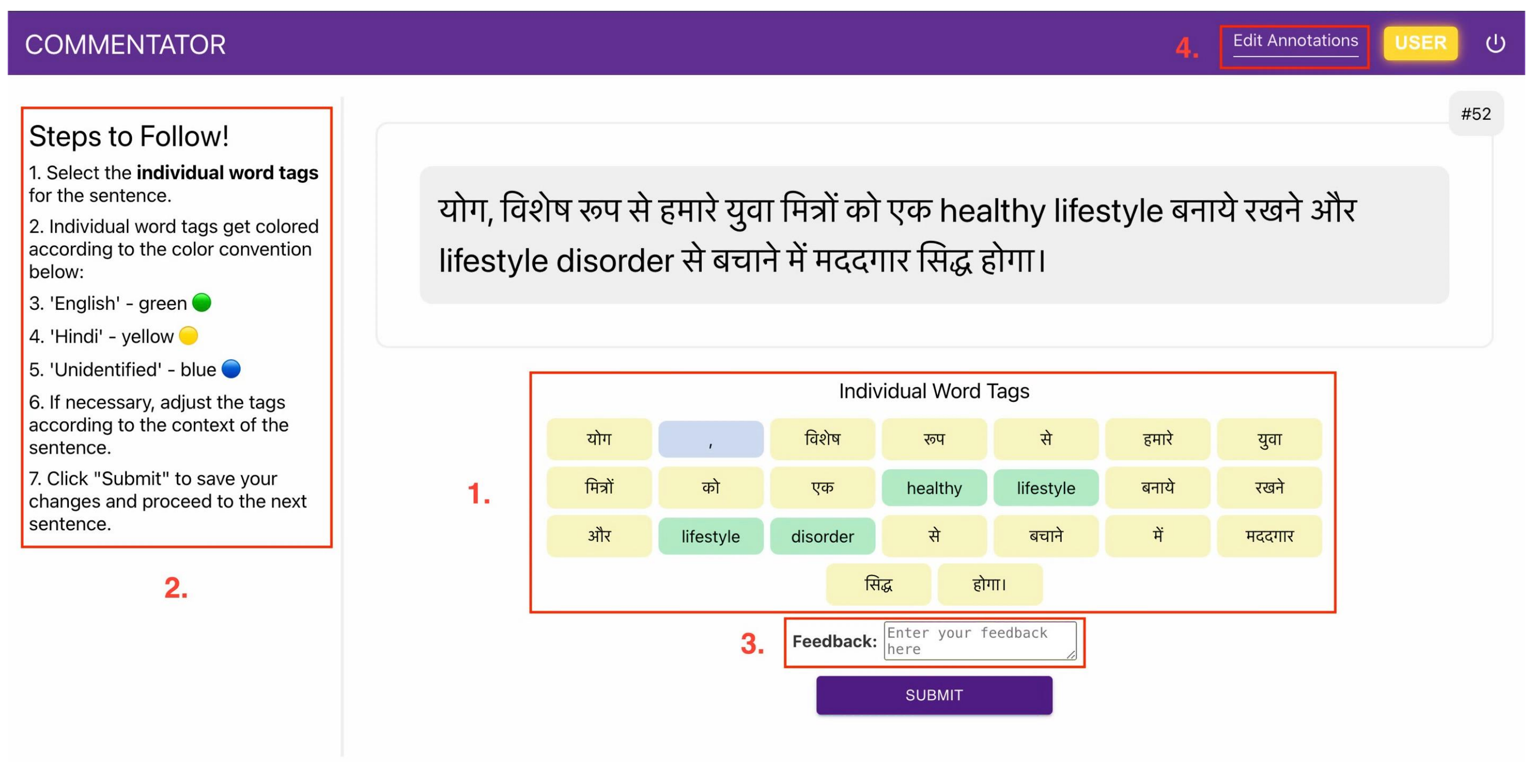
As shown in *Figure 1*, the **COMMENTATOR** architecture features a **ReactJS** client with an *Annotator panel* for user actions and an *Admin Panel* for data management. A **Flask** server connects to **MongoDB**, streamlining annotation with API calls.


1. Choose File No file chosen SUBMIT

2. ALL Enter CMI Threshold DOWNLOAD CSV

3. user1 user2 Enter Kappa Threshold DOWNLOAD COMPARISON CSV

Figure 2: The Admin panel of Commentator.



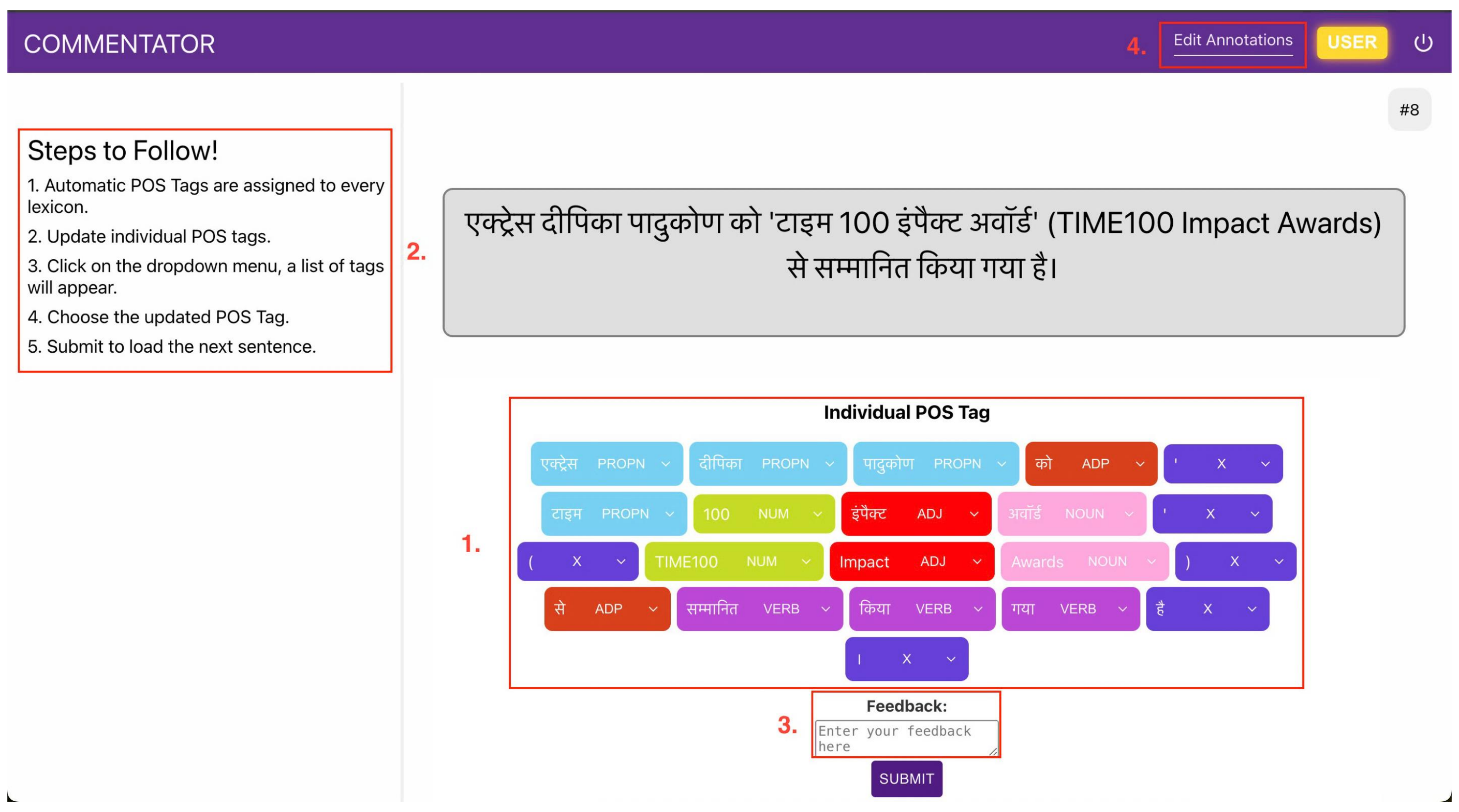
2. Steps to Follow!
1. Select the **individual word tags** for the sentence.
2. Individual word tags get colored according to the color convention below:
3. 'English' - green 🟢
4. 'Hindi' - yellow 🟡
5. 'Unidentified' - blue 🔵
6. If necessary, adjust the tags according to the context of the sentence.
7. Click "Submit" to save your changes and proceed to the next sentence.

योग, विशेष रूप से हमारे युवा मित्रों को एक healthy lifestyle बनाये रखने और lifestyle disorder से बचाने में मददगार सिद्ध होगा।

1. Individual Word Tags
योग , विशेष रूप से हमारे युवा
मित्रों को एक healthy lifestyle बनाये रखने
और lifestyle disorder से बचाने में मददगार
सिद्ध होगा।

3. Feedback: Enter your feedback here
SUBMIT

Figure 3: The Annotation panel for Token-level language identification (LID) task.



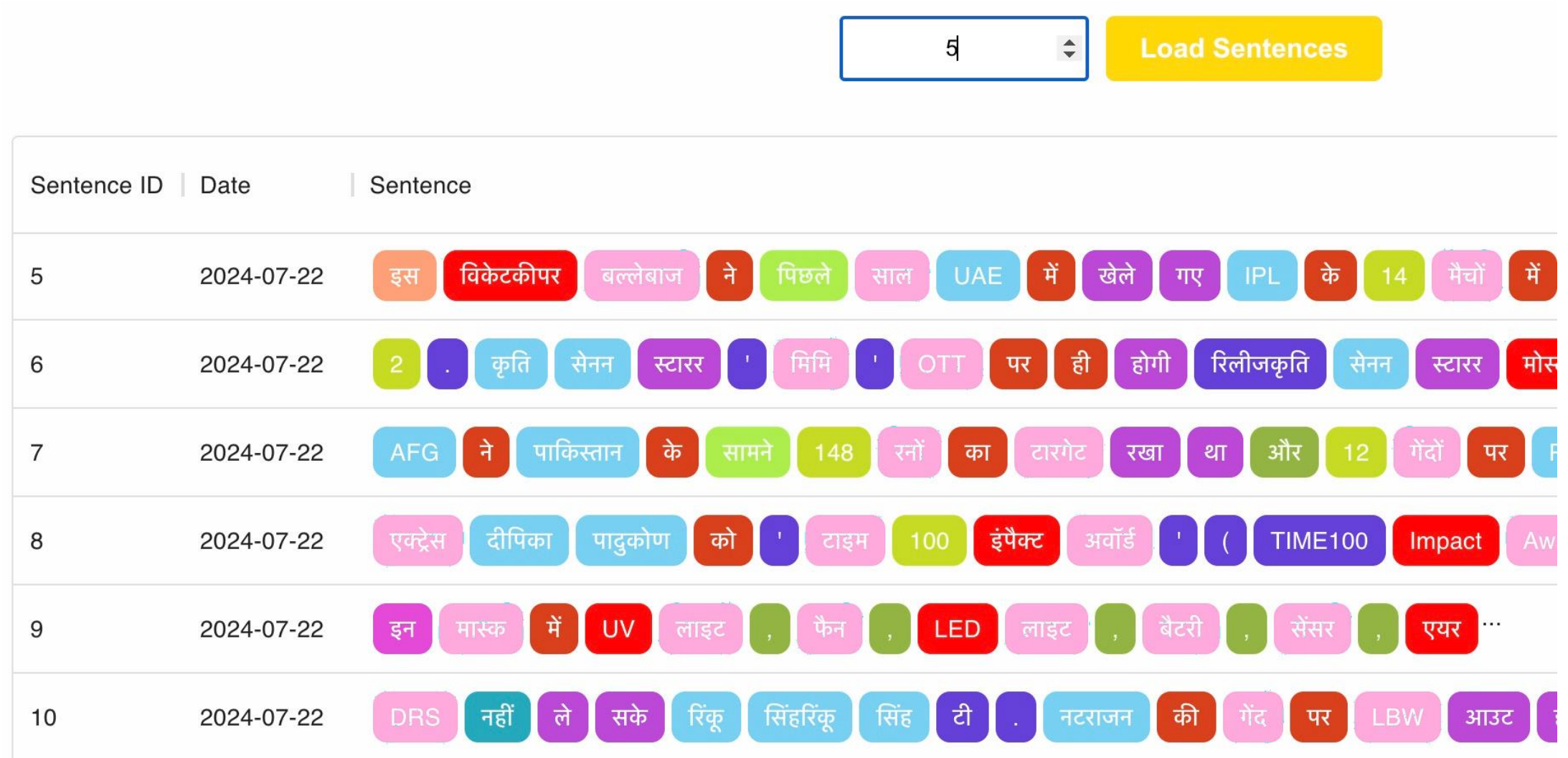
Steps to Follow!
1. Automatic POS Tags are assigned to every lexicon.
2. Update individual POS tags.
3. Click on the dropdown menu, a list of tags will appear.
4. Choose the updated POS Tag.
5. Submit to load the next sentence.

2. एक्ट्रेस दीपिका पादुकोण को 'टाइम 100 इंपैक्ट अवॉर्ड' (TIME100 Impact Awards) से सम्मानित किया गया है।

1. Individual POS Tag
एक्ट्रेस PROPN | दीपिका PROPN | पादुकोण PROPN | को ADP | ' X
टाइम PROPN | 100 NUM | इंपैक्ट ADJ | अवॉर्ड NOUN | ' X
(X | TIME100 NUM | Impact ADJ | Awards NOUN |) X
से ADP | सम्मानित VERB | किया VERB | गया VERB | है X
। X

3. Feedback: Enter your feedback here
SUBMIT

Figure 4: The Annotation panel for Token-level Part of speech Tagging (POS) task.


5 Load Sentences

Sentence ID	Date	Sentence
5	2024-07-22	इस विकेटकीपर बल्लेबाज ने पिछले साल UAE में खेले गए IPL के 14 मैचों में
6	2024-07-22	2 . कृति सेनन स्टारर ' मिमि ' OTT पर ही होगी रिलीजकृति सेनन स्टारर मोस
7	2024-07-22	AFG ने पाकिस्तान के सामने 148 रनों का टारगेट रखा था और 12 गेंदों पर F
8	2024-07-22	एक्ट्रेस दीपिका पादुकोण को ' टाइम 100 इंपैक्ट अवॉर्ड ' (TIME100 Impact Aw
9	2024-07-22	इन मास्क में UV लाइट , फैन , LED लाइट , बैटरी , सेंसर , एयर …
10	2024-07-22	DRS नहीं ले सके रिंकू सिंहरिंकू सिंह टी . नटराजन की गेंद पर LBW आउट

Figure 5: HISTORY AND EDIT page for POS task.

Evaluation

We conducted two studies to evaluate **COMMENTATOR**: the first *(Table 1)* **perceived capabilities** and the second *(Table 2)* demonstrated superior ***annotation speed***, highlighting its ***efficiency*** for multilingual NLP research.

Capabilities	*YEDDA[3]* *1 2 3*	*MarkUp[4]* *1 2 3*	*INCEpTION[5]* *1 2 3*	*UBIAI[6]* *1 2 3*	*GATE[7]* *1 2 3*	*BRAT[8]* *1 2 3*	*COMMENTATOR* *1 2 3*
Operational ease	✗ ✗ ✓	✓ ✓ ✗	✓ ✗ ✗	✗ ✓ ✓	✗ ✗ ✗	✓ ✓ ✗	✓ ✓ ✓
Less dependency requirements	✓ ✓ ✓	✓ ✓ ✓	✗ ✗ ✓	✗ ✓ ✓	✗ ✓ ✓	✓ ✓ ✗	✓ ✓ ✓
Low latency in API requests	✗ ✗ ✗	✗ ✓ ✗	✗ ✗ ✓	✓ ✗ ✗	✓ ✗ ✓	✗ ✗ ✗	✓ ✓ ✓
Admin Interface	✓ ✓ ✓	✓ ✓ ✓	✓ ✓ ✓	✓ ✓ ✓	✓ ✓ ✓	✗ ✗ ✗	✓ ✓ ✓
System recommendation	✓ ✓ ✗	✗ ✗ ✗	✓ ✓ ✗	✓ ✓ ✓	✓ ✗ ✗	✗ ✗ ✗	✓ ✓ ✓
Multiple user collaboration	✗ ✗ ✗	✗ ✓ ✗	✓ ✓ ✓	✓ ✓ ✓	✗ ✗ ✗	✓ ✓ ✓	✓ ✓ ✓
Annotation refinement and Feedback	✓ ✗ ✗	✗ ✓ ✓	✓ ✗ ✗	✓ ✓ ✓	✓ ✗ ✓	✓ ✓ ✓	✓ ✓ ✓
Post annotation analysis	✓ ✓ ✓	✓ ✓ ✓	✓ ✓ ✓	✓ ✓ ✓	✓ ✓ ✓	✗ ✗ ✗	✓ ✓ ✓

Table 1: Perceived capabilities by annotators. All annotators perceive all the eight capabilities in COMMENTATOR.

Frameworks	*LID*	*POS*
YEDDA	757.00 ± 62.27	1370.66 ± 81.24
MarkUp	1192.33 ± 172.77	1579.00 ± 68.86
INCEpTION	1040.66 ± 69.67	1714.66 ± 71.30
UBIAI	690.66 ± 79.43	748.33 ± 91.45
GATE	1118.33 ± 166.20	1579.00 ± 50.61
COMMENTATOR (ours)	**138.33 ± 24.60**	**337.66 ± 25.34**

Table 2: Average annotation time (mean ± SD) shows COMMENTATOR achieved **5x faster LID and 2x faster POS** annotations than the best baseline, UBIAI.

Conclusion

COMMENTATOR addresses annotation bias in ***Hindi-English*** code-mixed text annotation by integrating annotator ***feedback*** and calculating **IAA**, supporting three core NLP tasks, leading to a benchmark of over **100,000** instances.

References

[1] https://github.com/microsoft/LID-tool
[2] https://github.com/sagorbrur/codeswitch
[3] Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. 2018. Yedda: A lightweight collaborative text span annotation tool. ACL 2018, page 31.
[4] S Dobbie, H Strafford, WO Pickrell, B Fonferko-Shadrach, C Jones, A Akbari, S Thompson, and A Lacey. 2021. Markup: A web-based annotation tool powered by active learning. Frontiers in Digital Health, 3:598916–598916.
[5] Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations.
[6] UBIAI: NLP Annotation Tools - Automatic Text Annotation Tool." *UBIAI*, 2022, https://ubiai.tools/.
[7] Kalina Bontcheva, Hamish Cunningham, Ian Roberts, Angus Roberts, Valentin Tablan, Niraj Aswani, and Genevieve Gorrell. 2013. Gate teamware: a web-based, collaborative text annotation framework. Language Resources and Evaluation, 47:1007–1029.
[8] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for nlp-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 102–107, Avignon,