



Department of Computer Science, North Carolina State University *

{rajvee.sheth, heenaben.prajapati, himanshubeniwal, singh.mayank}@iitgn.ac.in, shubhnisar123@gmail.com

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

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 **LID¹ task** and **2x** for **POS² task** over the best baseline.

COMMENTATOR

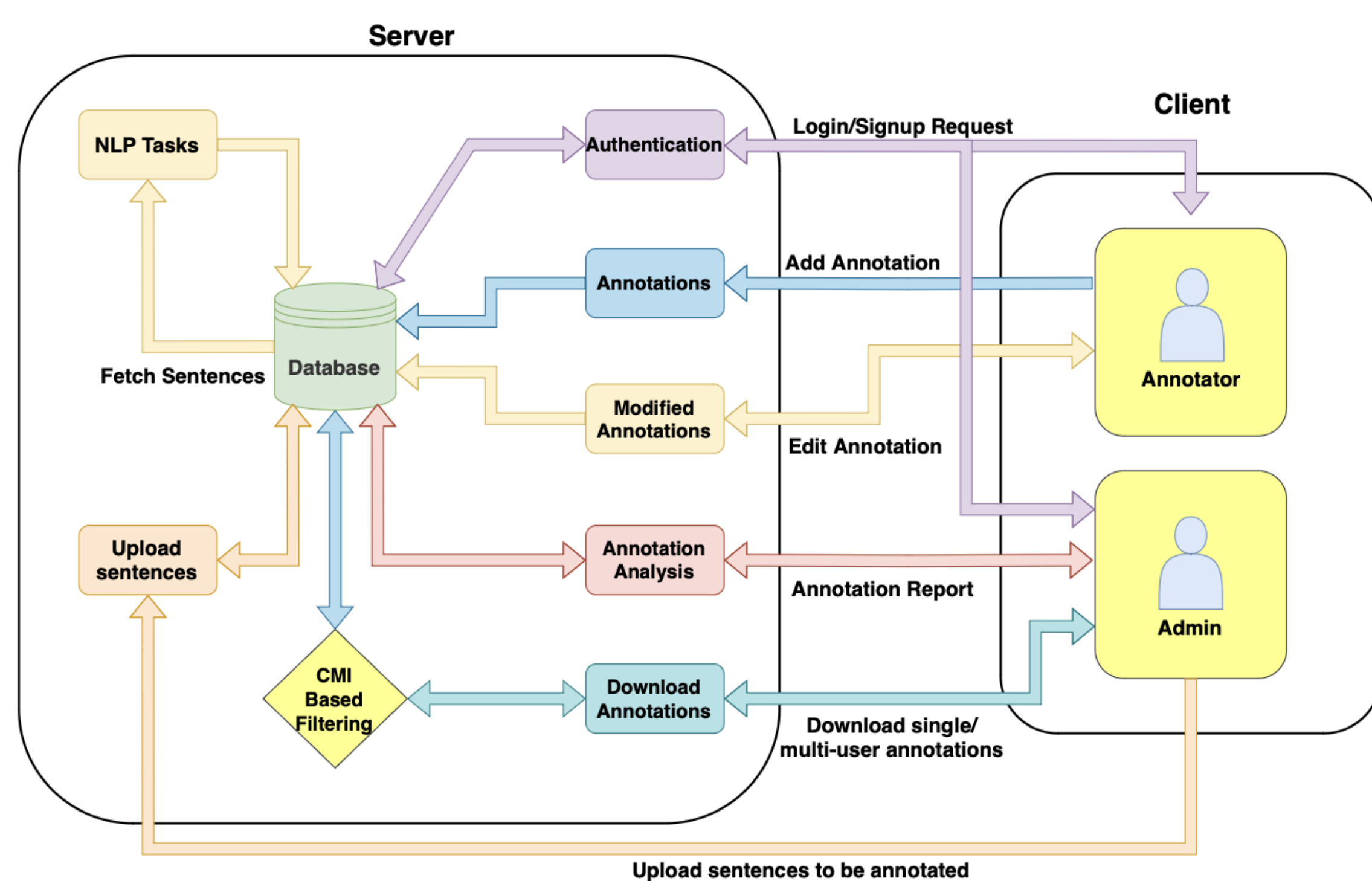


Figure 1: The proposed architecture of Commentator.

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.

COMMENTATOR

4.

Edit Annotations

USER

#52

Steps to Follow!

- Select the **individual word tags** for the sentence.
- Individual word tags get colored according to the color convention below:
- 'English' - green
- 'Hindi' - yellow
- 'Unidentified' - blue
- If necessary, adjust the tags according to the context of the sentence.
- Click "Submit" to save your changes and proceed to the next sentence.

2.

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

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.

COMMENTATOR

4. Edit Annotations
USER
⚙️

Steps to Follow!

- Automatic POS Tags are assigned to every lexicon.
- Update individual POS tags.
- Click on the dropdown menu, a list of tags will appear.
- Choose the updated POS Tag.
- Submit to load the next sentence.

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

1. Individual POS Tag

एकट्रेस 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 ▾	थिया VERB ▾	था VERB ▾	है X ▾
[X ▾				

Feedback:

3. Enter your feedback here.

SUBMIT

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

Sentence ID		Date	Sentence
5		2024-07-22	इस विकीट्रेडर बल्लेबाज ने विक्रमोत्तम भारत UAE में खेले गए TPL के 14 मैचों में
6		2024-07-22	ऊँची कुर्तियाँ सेना स्तर पर मित्रों OTAN पर ही होगी मिलजुबत सेना स्तर में
7		2024-07-22	AFG ने पाकिस्तान के सामने 148 रनों का झररोट रखा था और 12 गेंदों पर
8		2024-07-22	एकदम शीघ्रता मादुरोकन को 1 टाइम 100 इंचक आर्सेनल (TIME100 Impact
9		2024-07-22	गुरु बाबू से UV लाइट के बैंग LED लाइट के बेटों के बैंग एयर
10		2024-07-22	DRS नहीं ले सके सिर्फ सिंसिस्के सिंह टी नंदनराज 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	COMMENT ATOR 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.

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

Table 2: Average annotation time (mean \pm 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

<https://github.com/microsoft/LDTool>

<https://github.com/amazonsagebrush/codeswitch>

Jie Tang, Yue Zhang, Liwei Li, and Xingqun Li. 2018. Yekda: A lightweight collaborative text span annotation tool. ACL 2018, page 31.

Matt J. Dorr, David S. Stanford, Eyal Seifberg-Shadrach, C. James Alabi, S. Thompson, and A. Lacey. 2021. Markup: A web-based annotation tool powered by active learning. Frontiers in Digital Health, 3:589861–589816.

Xin-Christoph Kue, Michael Bugert, Boud Bouloosa, Richard Eckart de Castilho, and Iryna Geysewath. 2018. The INCEPTon platform: Machine-assisted and knowledge-oriented interactive annotation. In Proceedings of the 27th International Conference on Computational Linguistics, System Demonstrations.

IBM. 2021. Automatic Text Annotation Tool | IBM®. 2021. <https://c4liaib.net/>.

Kalina Bontcheva, Harriett Cunningham, Ian Roberts, Angus Roberts, Valerian Talari, Niraj Aswani, and Genevieve Gorell. 2013. Gate teamwork: a web-based, collaborative text annotation framework. Language Resources and Evaluation, 47:1007–1029.

Rafael Alvarez, Samy Pysalo, Goran Topić, Tomoko Ohira, Sophia Annadani, and Jun'ichi Tsuji. 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.