First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages (IndoNLP2025)

Abu Dhabi, United Arab Emirates
20 January 2025

ISBN: 979-8-3313-1369-2

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2025) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2026)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

Hindi Reading Comprehension: Do Large Language Models Exhibit Semantic Understanding?
Daisy Monika Lal, Paul Rayson and Mo El-Haj 1

Machine Translation and Transliteration for Indo-Aryan Languages: A Systematic Review
Sandun Sameera Perera and Deshan Koshala Sumanathilaka 11

BERTopic for Topic Modeling of Hindi Short Texts: A Comparative Study
Atharva Mutsaddi, Anvi Jamkhande, Aryan Shirish Thakre and Yashodhara Haribhakta 22

Evaluating Structural and Linguistic Quality in Urdu DRS Parsing and Generation through Bidirectional Evaluation
Muhammad Saad Amin, Luca Anselma and Alessandro Mazzei 33

Studying the Effect of Hindi Tokenizer Performance on Downstream Tasks
Rashi Goel and Fatiha Sadat 44

Adapting Multilingual LLMs to Low-Resource Languages using Continued Pre-training and Synthetic Corpus: A Case Study for Hindi LLMs
Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar and Eileen Long 50

OVQA: A Dataset for Visual Question Answering and Multimodal Research in Odia Language
Shantipriya Parida, Shashikanta Sahoo, Sambit Sekhar, Kalyanamalini Sahoo, Ketan Kotwal, Sonal Khosla, Satya Ranjan Dash, Aneesh Bose, Guneet Singh Kohli, Smruti Smita Lenka and Ondřej Bojar 58

Advancing Multilingual Speaker Identification and Verification for Indo-Aryan and Dravidian Languages
Braveenan Sritharan and Uthayasanker Thayasivam 67

Sentiment Analysis of Sinhala News Comments Using Transformers
Isuru Bandaranayake and Hakim Usoof 74

ExMute: A Context-Enriched Multimodal Dataset for Hateful Memes
Riddhiman Swanan Debnath, Nahian Beente Firuj, Abdul Wadud Shakib, Sadia Sultana and Md Saiful Islam 83

Studying the capabilities of Large Language Models in solving Combinatorics Problems posed in Hindi
Yash Kumar and Subhajit Roy 90

From Scarcity to Capability: Empowering Fake News Detection in Low-Resource Languages with LLMs
Hrithik Majumdar Shibu, Shrestha Datta, Md. Sumon Miah, Nasrullah Sami, Mahruba Sharmin Chowdhury and Md Saiful Islam 100

Enhancing Participatory Development Research in South Asia through LLM Agents System: An Empirically-Grounded Methodological Initiative from Field Evidence in Sri Lankan
Xinjie Zhao, Hao Wang, Shyaman Maduranga Sriwarnasinghe, Jiacheng Tang, Shiyun Wang, Sayaka Sugiyama and So Morikawa 108

Identifying Aggression and Offensive Language in Code-Mixed Tweets: A Multi-Task Transfer Learning Approach
Bharath Kancharla, Prabhjot Singh, Lohith Bhagavan Kancharla, Yashita Chama and Raksha Sharma 122

Team IndiDataMiner at IndoNLP 2025: Hindi Back Transliteration - Roman to Devanagari using LLaMa
Saurabh Kumar, Dhruvkumar Babubhai Kakadiya and Sanasam Ranbir Singh 129

IndoNLP 2025 Shared Task: Romanized Sinhala to Sinhala Reverse Transliteration Using BERT
Sandun Sameera Perera, Lahiru Prabhath Jayakodi, Deshan Koshala Sumanathilaka and Isuri Anuradha 135