Survey of Pseudonymization, Abstractive Summarization & Spell Checker for Hindi and Marathi

Rasika Ransing, Mohammed Amaan Dhamaskar, Ayush Rajpurohit, Amey Dhoke, Sanket Dalvi

Vidyalankar Institute of Technology
{rasika.ransing, mohammedamaan.dhamaskar, ayush.rajpurohit, amey.dhoke, sanket.dalvi}@vit.edu.in

Abstract

India's vast linguistic diversity presents unique challenges and opportunities for technological advancement, especially in the realm of Natural Language Processing (NLP). While there has been significant progress in NLP applications for widely spoken languages, the regional languages of India, such as Marathi and Hindi, remain underserved. Research in the field of NLP for Indian regional languages is at a formative stage and holds immense significance. The paper aims to build a platform which enables the user to use various features like text anonymization, abstractive text summarization and spell checking in English, Hindi and Marathi language. The aim of these tools is to serve enterprise and consumer clients who predominantly use Indian Regional Languages.

1 Introduction

In the exponential expansion of digital communications and India's progressing toward the "Digital India Campaign", the necessity for tools and services that can fulfil the need of people to understand the texts and documents in Indian regional languages has expanded. While existing text processing tools effectively support prominent languages such as English, the development of comparable services for Indian regional languages remains significantly underdeveloped, hindering progress in these domains.

This platform will be designed to empower services in regional languages like Hindi and Marathi which are two of the most spoken languages in India by leveraging advanced NLP techniques. The platform will incorporate following core functionalities:

1. Text anonymization: In today's highly connected world, there are vast amounts of data that is generated every minute. These large volumes of data must be collected, stored and indexed before it is used for various downstream tasks. However, most of this data is strife with Personally Identifiable Information (PII). PII can be any information like Name, Gender, Organization, etc which can be used independently or in combination with other PII to identify an individual or an organization. The exposure of PII can have severe consequences including identity theft, financial fraud, and reputational damage. The potential for misuse of PII is especially concerning in contexts where sensitive information, such as medical records, financial data, or biometric details, is involved. To solve this issue, all data points should be anonymized before publishing. Anonymization simply means removing this sensitive information. This step is crucial to protect individuals' privacy and mitigate the risks associated with the PII exposure.

2. Text summarization: One of today's major challenges is managing information overload so users can quickly grasp key ideas from large documents. Summaries enhance the efficiency of information sharing, enabling people to understand

essential points regardless of their literacy skills. By offering a quick overview, summaries help users decide the relevance of a document. This is particularly valuable in fields like research, legal analysis, and data extraction, where finding the right information is crucial. In NLP, summarization also supports other features like text analysis, translation, and sentiment analysis.

3. Spell correction: With the increase in various tools and platforms, there is a need for people to effectively and correctly present their ideas. For any writing tool having a basic spell checker is essential as it improves accuracy and efficiency of the entire process. For non-native speakers or learners having a good spell checker which supports various languages improves the overall experience of writing content in any desired language. This paper talks about various techniques used for implementing multilingual spell checkers.

The primary objective of this platform is to facilitate communications in these languages to progress towards the digital India future in hand with the regional languages. By fulfilling the demands for text processing in these languages the platform will seamlessly facilitate the services in English along with Hindi and Marathi.

This paper explores the implementation of existing systems and technologies along with approaches and methodologies that will be followed to design the platform. The following sections will detail the literature survey and proposed system along with the potential impact of the platform on multilingual services.

2 Literature Survey

2.1 Text Anonymization

Anonymization is a data protection technique that involves transforming or removing Personally Identifiable Information (PII) from data to make it unidentifiable. Anonymization has two stages: first is the detection of sensitive information from a given text and second is replacing the detected sensitive information with placeholder or random values. For the first stage, there are various techniques to detect PII like Rule Based tagging, dictionary lookups, part-of-speech (POS) tagging and named entity recognition (NER), but the latter

two offer the best results, especially while handling out of vocab terms. POS tagging involves tokenizing sentences into words or tokens and assigning a grammatical class to each token. When POS tagging was introduced, it used classical techniques like Maximum Entropy Modelling [1]; however with the advancements in natural language processing, it evolved to using various modern machine learning models like LSTMs, CRFs and neural networks [2]. While part-of-speech was quite effective during its advent, it was later overshadowed by the advancements in named entity recognition. The term 'named entity' was first introduced at the sixth Message Understanding Conference [3], where the task of such a system was to detect classic entities like persons, locations and organizations in text. Since then, the task of named entity recognition has evolved to also include fine-grained NER to detect subcategories in entities [4] and nested NER to detect nested entities [5].

[6] and [7] compare the use of CRFs, LSTMs, Language Models and Transformer models for Named Entity Recognition (NER). Both of these works highlight the efficiency of transformer-based NER models, which consistently outperform traditional approaches. While these results are remarkable, much like most of the work in NLP, they focus only on the English language.

Across the past works on the Hindi language, FIRE 2014 [8] dataset (NER dataset for English, Hindi, Tamil and Malayalam) and WikiANN [9] dataset (NER dataset of 10,000 sentences for 282 languages) have garnered a lot of interest and led to multiple adaptations of different pretrained models for the task of Hindi NER. However, the first true gold standard dataset for Hindi NER was the HiNER [10] dataset and corresponding NER models released in 2022. It comprises over 100,000 manually annotated sentences and 11 entity categories.

Similarly for Marathi, until recent years, the only available NER corpus were the WikiANN dataset (included Marathi language) and the IIT Bombay Marathi NER Corpus [11], both of which were limited in dataset size and the number of entities. A breakthrough in the field of Marathi NER was the L3Cube-MahaNER [12] dataset and BERT models, which comprises 25,000 manually annotated sentences and 7 entity categories. Another noteworthy mention is the Naamapadam [13] dataset and IndicNER models which provide

a NER corpus for 11 Indian languages including Hindi and Marathi. We aim to further expand on this previous work and develop a single model which can perform NER for Hindi and Marathi languages.

The second stage involves the replacement of the detected entities with certain placeholder values. There are various anonymization techniques like Suppression (replaces each character of detected entity with '*'), Permutation (changes sequence of characters), Hashing (generates unique hash for each detected entity), Generalization (replaces detected entity with a generalized term (e.g. - Zip Code is replaced by the city name)) and Pseudonymization (replaces the detected PII with a semantically similar placeholder).

While all these techniques provide privacy and prevent the exposure of PII, Pseudonymization is not only the most sophisticated, but also the most effective in preserving the semantic usefulness of the context. This makes it the most effective technique in sanitizing text and datasets for different NLP tasks. The generation of pseudonyms for replacement is a complex task, and is driven by various characteristics like token length, variation in token styles and sub-categories within entities.

Due to these complexities, pseudonym generation is an active area of research in the NLP space, with various approaches like Dictionary lookups, rule-based matching and the use of generative models. [14] uses Dictionary lookup for Pseudonymization. The dictionary was prepared by curating data points from WikiData and replacements were chosen at random from the dictionary. [15] also compared anonymization across methods involving LLMs to generate pseudonyms and Seq-2Seq model. [16] uses NER for detection and fine-tuned BERT models for pseudonym generation. While these works are quite comprehensive in their research, most of this research is limited to the English language, with little work being done for Indian regional languages. Our research aims to implement these techniques for Hindi and Marathi language and publish a series of Pseudonym generation models for the same.

2.2 Abstractive Text Summarization

Early approaches to text summarization relied predominantly on extractive approaches, but with the significant advancement in the field of Natural Language Processing the opportunities to perform the task of summarizing via an advanced or more efficient approach is possible. Abstractive text summarization, a more efficient approach of extractive summarization is possible using the neural network models to generate more coherent and human like summaries.

In [17] the authors investigated several automatic text summarization (ATS) systems, both abstractive and extractive methods. They highlight the importance of artificial intelligence, particularly the use of the sequence-to-sequence (seq2seq) model in extracting of central ideas. These models are different from the extractive techniques in way that they re-word content to express main ideas rather than simply taking out sentences that are important. The journal also examines recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and transformer-based models as the standard architectures used to develop summarization methods. [18] proposes Hindi text summarization with a challenge and introduces IndicBART, a transformer model that is pre-trained on a multilingual corpus and fine-tuned on Hindi. Through transfer learning, IndicBART shares information from the English language to be effective in Hindi, gaining a ROUGE-1 F1 score of 0.544 for abstractive summaries, thus already promising despite little data.

[19] and [20] focus on abstractive summarization for Marathi. Even though transformer-based models improved summary quality significantly, they do not reach human-like results. Paper [19] reports attention-based and stacked LSTM Seq2Seq models using stop-word and rare-word lists for Hindi and Marathi with ROUGE scores of 0.61 (recall) and 0.625 (precision) whereas [20] presents a Deep Belief Network and Decision Tree based approach that achieves 95.49% precision and 92.76% accuracy.

[21] provides a detailed overview of how the script, particularly Hindi, has evolved regarding abstractive text summarization. Like [19], the discussion revolves around the complexity of the script and sentence structure. The authors compare traditional rule-based systems with modern neural networks and explain why these models are effective while processing Hindi text. Detailed inspection of seq2seq models [22] shows that they can be used for Hindi, mainly due to the effectiveness of these models in handling complex sentences and in producing coherent summaries.

[17] rightly emphasizes the need for attention mechanisms for improving the quality of Hindi summaries. In fact, promising results have already been observed with the transformer-based BERT, GPT & T5 fine-tuned models for Hindi, following pre-training on large multilingual corpora.

[23] proposed a transliteration-based approach to improving Hindi text summarization. The authors recommend representing Hindi text in Romanized form before processing. This removes all word segmentation and character ambiguities. The results in fine-tuning with this method generated a 0.5510 ROUGE-1 score. The authors address the potential benefits of this method over classic techniques regarding both semantic coherence and fluency in adaptation. New possibilities of cross-lingual adaptation of NLP models arise with this approach as well.

[24] aims at creating a news summarization tool that includes many ML models, including BERT and TextRank, in combination with NLP techniques like Named Entity Recognition (NER). It is mainly extractive in nature, but this marks the first step towards abstractive techniques by improving the contextual knowledge developed in the machine and generating human-like summaries. The combination of NLP and ML marks the transition from traditional techniques towards advanced neural network-based techniques for making the extraction of the news summarization.

[25] describes in detail datasets for Indian languages, which are significant for training models in regional languages. The authors have pointed out a problem of such scarce annotated datasets for abstractive summarization systems. They offer a comparative analysis of these datasets based on size, diversity, and content. The analysis stands out to be crucial for understanding current progress in such diverse datasets. As stated in [19] and [20], the key role that transfer learning and pre-trained models play is distinguished to overcome possible limitations due to scarcity of large dataset for Indian regional languages.

[26] presents a Hindi text summarization method that is a combination of extractive and abstractive methods. This is done by using a two-step process: first by choosing important sentences through the feature positions, keyword frequency, and title similarity, and then, after removing the redundant contents, making a good summary of the text. The method would have been better if it was rather based on sentence selection and elimination, however, the authors imply that combining modern methods, e.g. neural networks or transformers, will cause a significant improvement.

2.3 Spell checker

One of the earliest approaches to spell checking involved dictionary-based techniques. These methods identify misspelled words by checking each word against a predefined dictionary. While this approach works well for standard languages with comprehensive lexicons, it becomes challenging when applied to regional languages such as Marathi and Hindi. These languages are underrepresented in digital corpora, making dictionary creation and maintenance difficult [27].

Another approach includes probabilistic and machine learning-based methods. N-gram models, for example, analyze word sequences to predict the most likely word combinations, which aids in detecting and correcting spelling errors within context. In languages like Hindi and Marathi, where context plays a significant role in meaning, these methods have shown promise. However, challenges remain in obtaining large enough datasets for these models to function effectively in non-English languages [27,28].

Another modern approach involves the use of edit distance algorithms, such as the Levenshtein distance, to quantify how different two strings are and suggest corrections based on the closest matching word. While edit distance methods work well for detecting simple typos, they struggle with more complex errors, particularly in languages with script and orthography distinct from English.

Hybrid methods that combine dictionary-based techniques with machine learning models have also been explored in recent years. These methods aim to leverage the strengths of both approaches by using machine learning to identify likely corrections while still ensuring that suggestions conform to the basic linguistic rules of the language. Such techniques are especially useful in handling challenges posed by the morphological complexity of Marathi and Hindi [27,28].

One major obstacle is the lack of extensive linguistic resources, such as annotated corpora and comprehensive dictionaries, for these languages. Unlike English, where vast amounts of data are available for training and testing, Indian languages suffer from a scarcity of high-quality linguistic datasets. This limits the performance of both rule-

based and machine learning-driven spell checkers, as they require large datasets to function effectively [29].

[29] explores the intricacies of creating a spell-checking system for the Shahmukhi script, a less commonly studied script used in the Punjabi language. The design considerations, such as the identification of specific phonetic variations, typographical errors, and the integration of linguistic rules, offer valuable insights that can be extended to other regional languages like Marathi and Hindi. The method presented uses rule-based and statistical approaches to develop a comprehensive spell checker that considers the linguistic characteristics of the target language. This combination of approaches ensures a higher rate of error detection and correction, which is crucial for regional languages with rich morphological structures [30].

In the context of multilingual NLP systems, the application of spell-checking algorithms must address not only the diverse character sets and scripts but also the nuances in grammar and vocabulary. The Shahmukhi spell checker highlights how leveraging specific linguistic features, such as affixes and phonetic patterns, can be pivotal for improving the spell-checking accuracy in non-Latin scripts [29]. The inclusion of a large and well-annotated corpus further enhances the system's performance, a strategy that is equally applicable to Marathi and Hindi. The use of dictionaries and corpora can greatly aid in developing spell checkers that are not only accurate but also contextually aware, reducing the likelihood of inappropriate corrections.

3   Proposed System

3.1   Text Anonymization

We aim to develop an anonymization engine for Indian Regional languages, using pseudonymization [15] technique, with a focus on Hindi and Marathi. This requires the use of a NER model for identification of PII and using custom techniques to implement pseudonymization. We plan to experiment with various generative models to generate pseudonyms for Hindi and Marathi.

The language specific NER models outlined in [10] and [12] appear to outperform IndicNER, as indicated by their respective evaluation techniques. However, there is a scarcity of such language specific NER models, due to a lack of a large

corpus. We aim to address this problem by training a multilingual NER model to handle both Hindi and Marathi language and compare our results with that of MahaNER and HiNER based models.
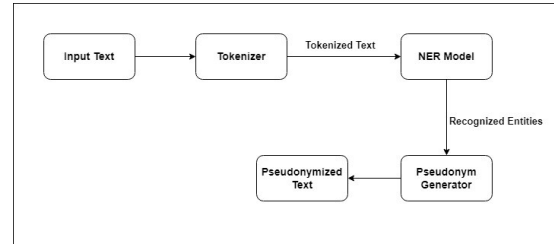

Figure 1: Proposed System for Text Anonymization

3.2   Abstractive Text Summarization

The task for abstractive text summarization for Hindi would be carried out using the *"Hindi Text Short and Large Summarization Corpus"* [31,32] dataset along with the Hindi language dataset provided by ILSUM.

For Marathi summarization we chose to translate a dataset provided by *"CNN-DailyMail News Text Summarization"* [33] with number of instances over 300k into Marathi language out of which we have already complete the translation of approximately 60k instances at the time of writing this survey paper.
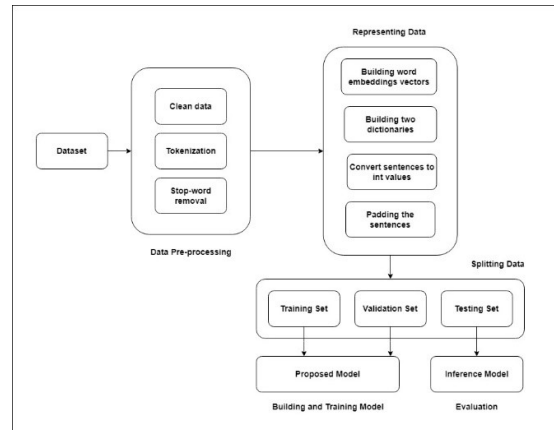

Figure 2 Abstractive Text Summarization Proposed System

The model we chose to use are:

3.2.1   BART based model

IndicBART provided by AI4Bharat based on BART which is a denoising autoencoder that maps a corrupted document to the original document through a sequence-to-sequence model with a

bidirectional encoder over corrupted text and a left-to-right autoregressive decoder [18].

Fine-tuning Facebook's BART-large-cnn model involves adapting the pre-trained model, optimized for English text summarization, to work with Marathi and Hindi datasets allowing benefits of pre-existing fine-tuning on CNN dataset.

3.2.2 LSTM-based models

The Time Distributed Stacked LSTM [19] model involves stacking two LSTM layers in the encoder making the model effective for capturing complex patterns in text data and hence making it useful for summarizing tasks. The first LSTM layer processes the input sequence, and its output is fed into the second LSTM layer.

The Attention mechanism is a significant enhancement to LSTM models, allowing the network to focus on specific parts of the input sequence. In the **Attention-Based Stacked LSTM** model [19], three LSTM layers are stacked together. This is achieved by using the attention helps the model decide which words in the input sequence are most relevant to the current word being generated in the summary.

3.2.3 Deep Recurrent Neural Networks (DRNN)

Deep Recurrent Neural Networks (DRNN) are an extension of traditional RNNs with multiple hidden layers by stacking RNN layers, enabling the network to capture more complex patterns in the data [21] and overcome limitations of shallow RNNs allows model to capture intricate patterns in text .

3.3 Spell checker

Our approach for Spell checker includes four main stages: Error Detection, Generating Candidate Suggestion, Ranking Suggestions and Replacing Incorrect word with correct word.

3.3.1 Error Detection:

We keep our error-search strictly to non-words errors; for every token in a sentence, we check for its occurrence in the dictionary.

3.3.2 Generating Candidate Suggestion:

Given an unknown token, we generated a list of all known words within an edit distance of 2, calling them candidate suggestions. Two intuitive approaches for generating suggestions that work reasonably well on smaller datasets are: checking the edit distance of the misspelled word against all words in the dictionary faces challenge with large corpora, and generating a list of all words within an

edit distance of 2 from the incorrect spelling faces challenges with longer words.

3.3.3 Ranking Suggestions:

We use Minimum edit distance to rank the suggestions. Minimum edit distance measures how many operations are required to change one word into another. These operations can include:

- **Insertion**: Adding a character.
- **Deletion**: Removing a character.
- **Substitution**: Replacing one character with another.
- **Transposition**: Swapping two adjacent characters.

By calculating the edit distance for all possible correct words, we can rank them based on their distance from the incorrect word.

3.3.4 Replacing Incorrect word with Correct word:

At the end we compare the result from n-grams and the top-k result from suggested correct words and if there is a match in those two results, the matched word is replaced with the incorrect word. If there is no match in those two results, the top ranked suggested result is replaced with the incorrect word.
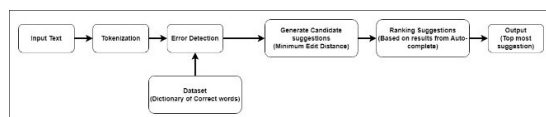

Figure 3: Spell Checker Proposed System

4 Conclusion

Recent advancement in the field of Natural Language Processing (NLP) have shown remarkable progress in applications namely text summarization, spell checking and data anonymization. But the development of equivalent tools in Hindi and Marathi is relatively nascent. While some research exists, a unified platform integrating various NLP tasks is still lacking.

The research highlights the potential of developing a user-friendly platform for text summarization, spell checking, anonymization and sentiment analysis especially for Marathi and Hindi language which significantly contribute to digital inclusion. Our proposed system focuses on addressing challenges like data scarcity and language complexity by implementing state-of-the-art AI and ML technologies by combining available resources with our own datasets.

References

[1] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71.

[2] Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

[3] Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A Brief History. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

[4] Antonia Höfer and Mina Mottahedin. 2023. Minanto at SemEval-2023 Task 2: Fine-tuning XLM-RoBERTa for Named Entity Recognition on English Data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1127–1130, Toronto, Canada. Association for Computational Linguistics.

[5] Girish Keshav Palshikar,2012. Techniques for named entity recognition: A Survey. *In Collaboration and the Semantic Web: Social Networks, Knowledge Networks, and Knowledge Resources* (pp. 191-217). IGI Global.

[6] Dimitris Asimopoulos, Ilias Siniosoglou, Vasileios Argyriou, Sotirios K. Goudos, Konstantinos E. Psannis, Nikoleta Karditsioti, Theocharis Saoulidis and Panagiotis Sarigiannidis. 2024. Evaluating the Efficacy of AI Techniques in Textual Anonymization: A Comparative Study. *arXiv preprint arXiv:2405.06709*.

[7] Dimitris Asimopoulos, Ilias Siniosoglou, Vasileios Argyriou, Thomai Karamitsou, Eleftherios Fountoukidis, Sotirios K. Goudos, Ioannis D. Moscholios, Konstantinos E. Psannis and Panagiotis Sarigiannidis. 2024. Benchmarking Advanced Text Anonymisation Methods: A Comparative Study on Novel and Traditional Approaches. *arXiv preprint arXiv:2404.14465*.

[8] N. Abinaya, Neethu John, Barathi H. B. Ganesh, Anand M. Kumar, and K. P. Soman. 2014. AMRITA_CEN@FIRE-2014: Named Entity Recognition for Indian Languages using Rich Features. In *Proceedings of the 6th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE '14)*. Association for Computing Machinery, New York, NY, USA, 103–111.

[9] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

[10] Rudra Murthy, Pallab Bhattacharjee, Rahul Sharnagat, Jyotsana Khatri, Diptesh Kanojia, and Pushpak Bhattacharyya. 2022. HiNER: A large Hindi Named Entity Recognition Dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4467–4476, Marseille, France. European Language Resources Association.

[11] Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2018. Judicious Selection of Training Data in Assisting Language for Multilingual Neural NER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–406, Melbourne, Australia. Association for Computational Linguistics.

[12] Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade, and Raviraj Joshi. 2022. L3Cube-MahaNER: A Marathi Named Entity Recognition Dataset and BERT models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 29–34, Marseille, France. European Language Resources Association.

[13] Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456, Toronto, Canada. Association for Computational Linguistics.

[14] Nuno Mamede, Jorge Baptista, Francisco Dias. 2016. Automated anonymization of text documents. In *IEEE Congress on Evolutionary Computation (CEC)*, Vancouver, BC, Canada, 2016, pp. 1287-1294.

[15] Oleksandr Yermilov, Vipul Raheja, and Artem Chernodub. 2023. Privacy- and Utility-Preserving NLP with Anonymized data: A case study of Pseudonymization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 232–241, Toronto, Canada. Association for Computational Linguistics.

[16] Ying Mo and Zhoujun Li. 2024. Incorporating Entity Type-Aware and Word–Word Relation-Aware Attention in Generative Named Entity Recognition. *Electronics*. 2024; 13(7):1407.

[17] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan and M. M. Kabir. 2021. A Survey of Automatic Text Summarization: Progress, Process and Challenges. IEEE Access, vol. 9, pp. 156043-156070.

[18] Arjit Agarwal, Soham Naik, and Sheetal Sonawane. 2022. Abstractive Text Summarization for Hindi Language using IndicBART

[19] R. Karmakar, K. Nirantar, P. Kurunkar, P. Hiremath and D. Chaudhari. 2021. Indian Regional Language Abstractive Text Summarization using Attention-based LSTM Neural Network. International Conference on Intelligent Technologies (CONIT), Hubli, India. pp. 1-8,

[20] Kale, S., Mahalle, P., Kachhoria, R., Kumar, S., Chaudhari, P., & Patil, V. 2023. Marathi text summarization through NLP and deep learning mechanism. *Journal of Autonomous Intelligence*. [S.I.], v. 6, n. 3.

[21] Aparna Madhukar Mete. 2024. A Comprehensive Survey on Abstractive Text Summarization of Devanagari Script Based Hindi Language. International Journal of Intelligent Systems and Applications in Engineering, 12(3), 3604–3620.

[22] Namrata Kumari, Pardeep Singh. 2022. Hindi Text Summarization using Sequence to Sequence Neural Network. PREPRINT (Version 1) available at Research Square.

[23] Jeetendra Kumar, Shashi Shekhar, Rashmi Gupta. 2024. Hindi Abstractive Text Summarization using Transliteration with Pre-trained Model.

[24] Ranjeeta Krishna, Vaishnavi Kahar, Atul Prasad Sharma, Divya Mishra. 2022. A Research Based Project on News Summarizer or News Summary Generator.

[25] Shangun Sinha, Girish Nath Jha. 2022. An Overview of Indian Language Datasets used for Text Summarization.

[26] Mili Supreet, Kanu Goel, Madhuri Gupta. 2020. Automatic Hindi Text Summarization Using Selection and Elimination Approach. International Journal of Engineering Applied Sciences and Technology, Vol. 5, Issue 4, ISSN No. 2455-2143.

[27] Sumreet, K. R., & Charanjiv, S. S. (2014). Study of Spell-Checking Techniques and Available Spell Checkers in Regional Languages: A Survey. *International Journal for Technological Research in Engineering, 2, 148-151.*

[28] Gupta, Prabhakar. (2020). A Context-Sensitive Real-Time Spell Checker with Language Adaptability. 116-122. 10.1109/ICSC.2020.00023.

[29] Dhanju, Kawarbir & Lehal, Gurpreet & Saini, Tejinder & Kaur, Arshdeep. (2015). Design and Implementation of Shahmukhi Spell Checker. *Indian Journal of Science and Technology. 8. 10.17485/ijst/2015/v8i27/83917.*

[30] P H, Hema & C, Sunitha. (2016). Malayalam Spell Checker Using N-Gram Method. 10.1007/978-81-322-2734-2_23.

[31] S. Satapara, B. Modha, S. Modha, P. Mehta. 2022. Findings of the first shared task on indian language summarization (ilsum): Approaches, challenges and the path ahead. Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, CEUR Workshop Proceedings, CEUR-WS.org, 2022.

[32] Shrey Satapara, Bhavan Modha, Sandip Modha, and Parth Mehta. 2023. FIRE 2022 ILSUM Track: Indian Language Summarization. In Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE '22). Association for Computing Machinery, New York, NY, USA, 8–11.

[33] Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, Bing Xiang. 2016. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond.