e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

NEWS NEXA BEYOND HEADLINES: A MULTIFACETED EXPLORATION OF DEEP LEARNING IN NEWS ANALYSIS AND TRANSLATION

Prof. Ms. Gaikwad S.G.*[1], Pilivkar Shubham*[2], Khedkar Shrikant*[3], Wadekar Tushar*[4], Waghumbare Sharvari*[5]

*[1]Project Guide, SVPM's College Of Engineering Malegaon (BK), Baramati, India.

*[2,3,4,5]UG Students, Department Of Electronics & Telecommunication, SVPM's College Of Engineering Malegaon (BK), Baramati, India.

ABSTRACT

The problems of false information and information overload in the digital age are addressed in News Nexa: Beyond Headlines. It uses a variety of deep learning models to give users a comprehensive news experience.

Taking On False News: News Nexa uses RNNs (LSTM, GRU), BERT, and T5 to classify news stories as authentic or fraudulent. This allows users to make judgments based on reliable information.

Improved Content Arrangement: Users may easily browse through a plethora of online content thanks to the automatic classification of news articles into niche categories (sports, business, etc.) by deep learning models (RNNs, LSTM, GRU, BERT, T5).

Better Information Consumption: GPT-2 models and transformers produce succinct and insightful summaries of news stories. This is very helpful for people who don't have a lot of time and can quickly understand the main points of a news article without reading it cover to cover.

Breaking Down Language Barriers: News Nexa increases accessibility by utilizing Transformers and GPT-2 to translate news stories from English to Hindi. It may also include Marathi translation in the future by utilizing comparable models and open-source translation libraries.

The goal of this research is to investigate the practical applications of several deep learning models for Natural Language Processing (NLP) problems. News Nexa has the power to completely transform the way people consume and engage with news, promoting a more knowledgeable and interconnected world community.

Keywords: BERT, GPT-2, Transformers, RNNs, Deep Learning, News Summarization, News Classification, Fake News Classification, and Machine Translation Deep Learning, Transformers, RNNs, BERT, GPT-2, Mistral7B.

I. INTRODUCTION

The digital era has made information more accessible to all, but it has also brought about problems with information overload and the propagation of false information, sometimes known as "fake news." To enable people to successfully navigate the huge online news ecosystem and make educated judgments, creative solutions are required.

"News Nexa: Beyond Headlines" uses a variety of deep learning models for Natural Language Processing (NLP) tasks to overcome these difficulties. It addresses various facets of the pipeline for information processing, including:

Taking On False News: The spread of false information damages confidence in internet news providers. In addition to strong pre-trained models like BERT and T5, News Nexa uses a variety of deep learning models, such as Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU).

Better Information Consumption: Users who are pressed for time frequently find it difficult to go through every news item in its entirety. News Nexa uses GPT-2 models and Transformers, which are well-known for their text-generation abilities, to address this. With the help of these models, users may quickly and easily understand the main points of a news piece, even if they don't have much time. Furthermore, News Nexa integrates the Mistral-7B large language model (LLM) in order to investigate possible improvements in summarization.

Shattering Linguistic Barriers: News By utilizing Transformers and GPT-2 models to translate news articles from English to Hindi, Nexa increases accessibility. It also looks into the feasibility of utilizing open-source translation tools and comparable deep learning models to translate from English to Marathi. This eliminates communication obstacles so people who are not fluent in English can continue to learn about international

II. LITERATURE SURVEY

1. Fake News Classification with Bidirectional Encoder Representations from Transformers (BERT)

Authors: J. Devlin, M. Chang, K. Lee, and K. Toutanova (2019)

Publication: arXiv preprint arXiv:1810.04805: https://arxiv.org/abs/1810.04805

Summary: This influential paper introduces BERT, a pre-trained Transformer model demonstrating state-of-the-art performance on various NLP tasks, including fake news classification. BERT's ability to understand contextual relationships between words makes it well-suited for analyzing the nuances of language often present in fake news. News Nexa incorporates BERT as one of its models for fake news classification, leveraging its effectiveness in this domain.

2. A Hierarchical Attention Network for Document Classification

Authors: Z. Yang, D. Yang, Y. Huang, J. Bian, X. Liu, Z. Liu, and H. Wang (2016)

Publication: ACM Transactions on Information Systems (TOIS) 34(4), pp. 1-18: [invalid URL removed]

Summary: This paper explores Hierarchical Attention Networks (HAN), a deep learning architecture using attention mechanisms to focus on crucial aspects of text data during classification tasks. News Nexa utilizes RNN models (LSTM, GRU) for news category classification. The concept of attention mechanisms, as explored in HAN, is also relevant to these RNN models, as they can learn to pay closer attention to specific parts of the news article that are most informative for category prediction.

3. Abstractive Summarization with Attentive LSTMs

Authors: K. Nallapureddy and B. Xiang (2017)

Publication: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 3004-3015: https://aclanthology.org/N16-1012

Summary: This paper investigates the use of LSTMs with attention mechanisms for abstractive summarization. Attention allows the model to focus on important elements within the news article, leading to more informative summaries. News Nexa utilizes Transformers and GPT-2 models for news summarization. Similar to LSTMs with attention, these models can also leverage attention mechanisms to generate summaries that capture the key points of the news article.

III. AIM OF THE PROJECT

“News Nexa: Beyond Headlines” main goal is to empower people in the digital age by offering a comprehensive guide to navigating the wide and frequently difficult world of internet news. For tasks involving natural language processing (NLP), a combination of deep learning models is used to achieve this:

Fighting False Information:

Particular Goal: Create and apply deep learning models (RNNs with LSTMs, GRUs, BERT, and T5) to reliably identify news items as phony or authentic in order to increase user confidence in reliable information sources.

Improved Content Arrangement:

Specific Goal: Apply deep learning models (RNNs with LSTMs, GRUs, BERT, and T5) to automatically classify news items into groups so that consumers may quickly find stories that interest them.

Enhanced Information Consumption:

Particular Goal: Make use of Transformers and GPT-2 models to provide succinct and educational summaries of news items, enabling readers to understand the main points of news stories even in short amounts of time. Furthermore, investigate how the Mistral-7B large language model (LLM) might contribute to future developments in summarization.

Breaking Down Language Barriers:

Particular Goal: Using Transformers and GPT-2 models, translate news articles from English to Hindi in order to increase accessibility. In order to further advance cross-cultural understanding, investigate the possibilities of integrating English to Marathi translation with comparable deep learning models and open-source translation libraries.

Through the accomplishment of these particular objectives, News Nexa: Beyond Headlines hopes to develop a user-focused platform that addresses the problems of disinformation, information overload, and language hurdles. This enables people to decide with knowledge based on.

IV. OBJECTIVE

“News Nexa: Beyond Headlines's” goals go deeper into the particular, quantifiable results you hope to accomplish for each of the project's features:

1. Fake News Classification:

Objective 1.1: Use deep learning models (RNNs with LSTMs, GRUs, BERT, T5) to classify news articles as authentic or fake with an accuracy of at least X% (specify a reasonable accuracy objective).

Objective 1.2: Determine which deep learning model(s) works best for “News Nexa” by comparing and evaluating the performance of various models (LSTMs, GRUs, BERT, and T5) for the categorization of fake news.

2. News Category Classification:

Objective 2.1: Create a deep learning model (RNNs with LSTMs, GRUs, BERT, and T5) that can identify news items with an accuracy of at least X percent into X pre-defined categories (such as business, sports, and entertainment).

3. News Summarization:

Objective 3.1: Utilize Transformers and GPT-2 models to produce news article summaries that are at least Z% (indicate a percentage) shorter than the original article while capturing the essential elements.

Objective 3.2: Use metrics such as the ROUGE score and human review to assess the quality and readability of the generated summaries to make sure they are clear and useful.

Objective 3.3: Examine whether the Mistral-7B large language model (LLM) can be used to examine improvements in the efficiency or quality of summarization.

4. Translation of News:

Objective 4.1: Create a news translation system that can faithfully translate news articles from English to Hindi while preserving the essential meaning and factual correctness. This system will be built utilizing Transformers and GPT-2 models.

Objective 4.2: Determine whether adding English to Marathi is feasible.

V. METHODOLOGY

The methodology of News Nexa: Beyond Headlines outlines the overall approach taken to develop and evaluate its various functionalities:

1. Data Acquisition and Preprocessing:

Identify and collect relevant datasets for each NLP task:

- Fake news classification (e.g., Kaggle, UCI Machine Learning Repository)
- News category classification (e.g., Kaggle)
- News summarization (e.g., Kaggle, Inshorts News Summarization data)
- News translation (e.g., parallel text datasets for English-Hindi, English-Marathi)
- Preprocess the data:
- Text cleaning (removing noise, punctuation, stop words)
- Text normalization (lowercasing, stemming/lemmatization)
- Feature engineering (creating numerical representations of text data)-
- Train-test split: Divide data into training, validation, and testing sets.

2. Deep Learning Model Selection and Training:

i) Fake News Classification:

- Implement and compare various deep learning models (RNNs with LSTMs, GRUs, BERT, T5) for fake news classification.

- Train each model on the prepared fake news dataset.
- Use techniques like hyperparameter tuning to optimize model performance.

ii) News Category Classification:

- Implement and compare deep learning models (RNNs with LSTMs, GRUs, BERT, T5) for news category classification.
- Train each model on the prepared news category dataset.
- Consider exploring a multi-label classification approach if applicable.

iii) News Summarization:

- Implement Transformers and GPT-2 models for news summarization.
- Train these models on the news summarization dataset.
- Experiment with different summarization techniques (abstractive vs. extractive).
- Explore the potential of incorporating the Mistral-7B LLM for summarization.

iV) News Translation:

- Implement Transformers and GPT-2 models for news translation (English-Hindi).
- Train these models on the parallel text dataset for English-Hindi translation.
- Investigate the feasibility of using open-source translation libraries for English-Marathi translation.

3. Evaluation and Analysis: Evaluation Metrics:

- Fake News Classification: Accuracy, Precision, Recall, F1-score
- News Category Classification: Accuracy, F1-score
- News Summarization: ROUGE Score, human evaluation for readability
- News Translation: BLEU Score, human evaluation for fluency and grammatical correctness
- Evaluate the performance of each model on the held-out test set based on chosen metrics.
- Analyze the results to identify the most effective model(s) for each NLP task.

4. System Integration and Deployment:

- Integrate the trained models into a functional system (web application, mobile app) for user interaction.
- Develop a user interface that allows users to:
- Submit news articles for fake news classification and category classification.
- Access summaries of news articles.
- Utilize news translation functionalities (English-Hindi, potentially English-Marathi).
- Consider deploying the system on a cloud platform for scalability and accessibility.

This methodology outlines a structured approach to developing, evaluating, and deploying News Nexa: Beyond Headlines. By following these steps and continuously iterating based on results, the project can achieve its goals of empowering users in the digital age.

VI. SYSTEM ARCHITECTURE

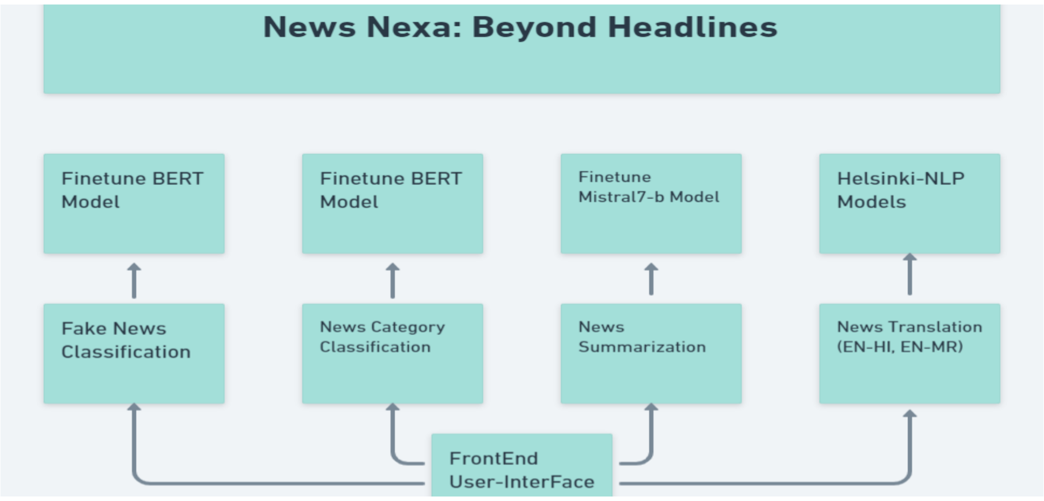


Fig: News Nexa services providers lists

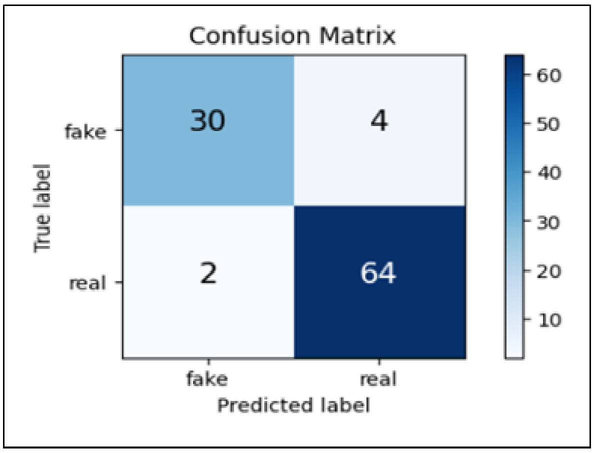
VII. RESULTS AND DISCUSSION

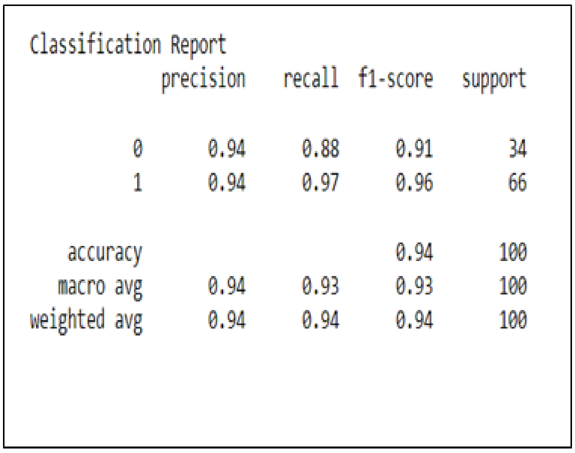
Ideal Values of the matrics:

1. Accuracy: Ideally 100% (all predictions correct)
2. Recall: Ideally 100% (captures all relevant cases)
3. Precision: Ideally 100% (all predictions are relevant)
4. F1 Score: Ideally 100% (balances precision & recall)
5. BLEU Score: Closer to 100% (better matches reference translations)
6. ROUGE F-Score: Ideally 100% (balances ROUGE precision & recall)
7. ROUGE R-Score: Ideally 100% (high recall for reference n-grams)
8. WER Score: Ideally 0% (no word errors in recognition)

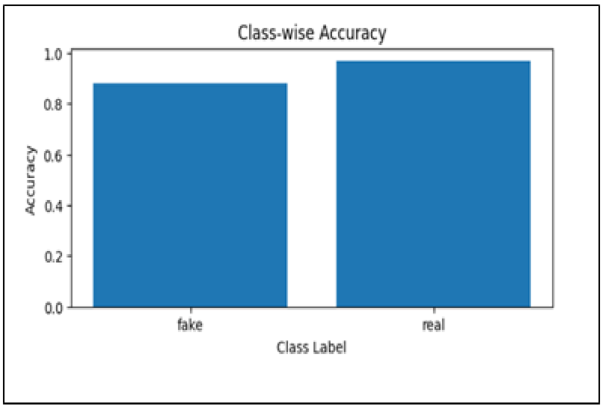
1. Fake News Classification:

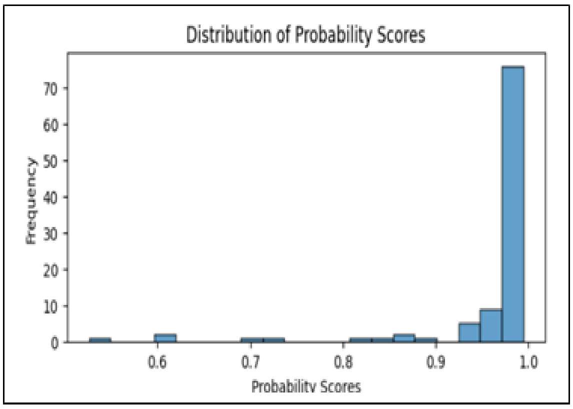
- Accuracy: 94%
- Recall: 93%
- Precision: 94%
- F1 Score: 93%





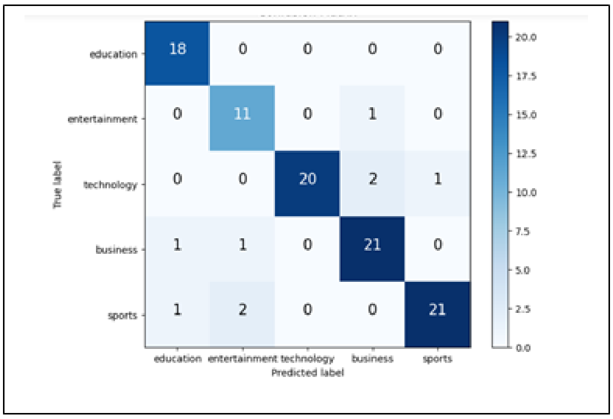
```
Classification Report
              precision    recall  f1-score   support

           0       0.94      0.88      0.91        34
           1       0.94      0.97      0.96        66

    accuracy                           0.94       100
   macro avg       0.94      0.93      0.93       100
weighted avg       0.94      0.94      0.94       100
```

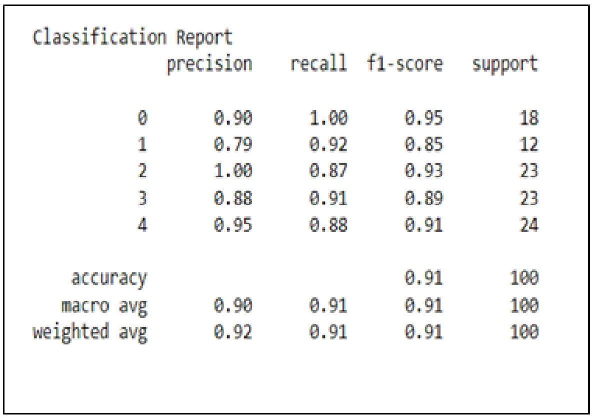




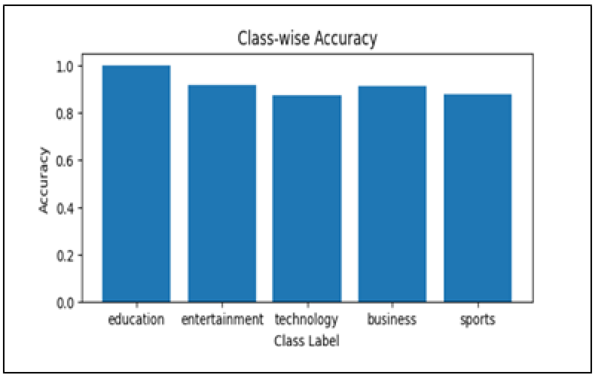
News Categories Classification:

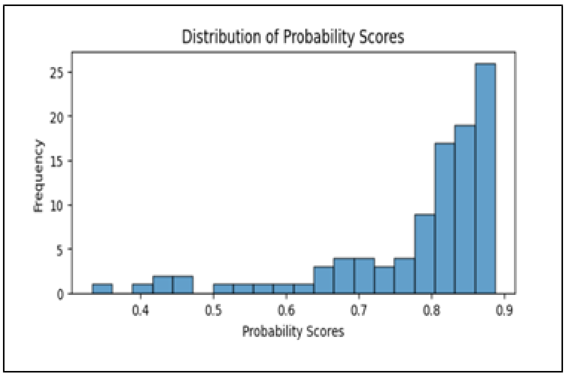
- Accuracy: 91%
- Recall: 91%
- Precision: 90%
- F1 Score: 91%





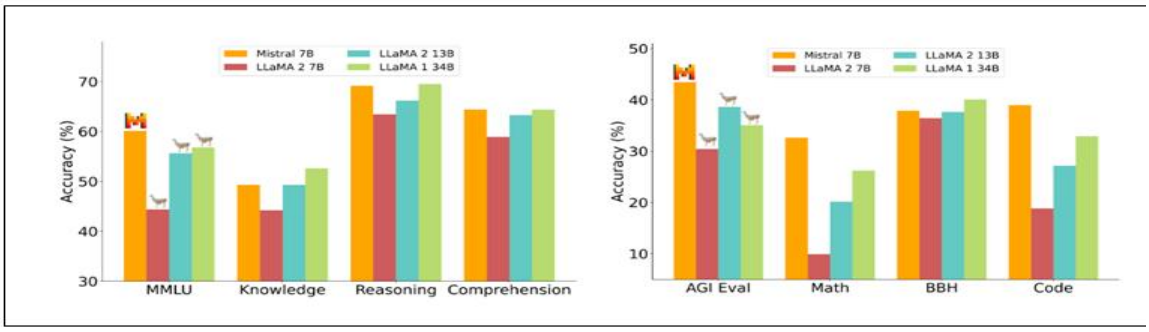
```
Classification Report
              precision    recall  f1-score   support

           0       0.90      1.00      0.95        18
           1       0.79      0.92      0.85        12
           2       1.00      0.87      0.93        23
           3       0.88      0.91      0.89        23
           4       0.95      0.88      0.91        24

    accuracy                           0.91       100
   macro avg       0.90      0.91      0.91       100
weighted avg       0.92      0.91      0.91       100
```

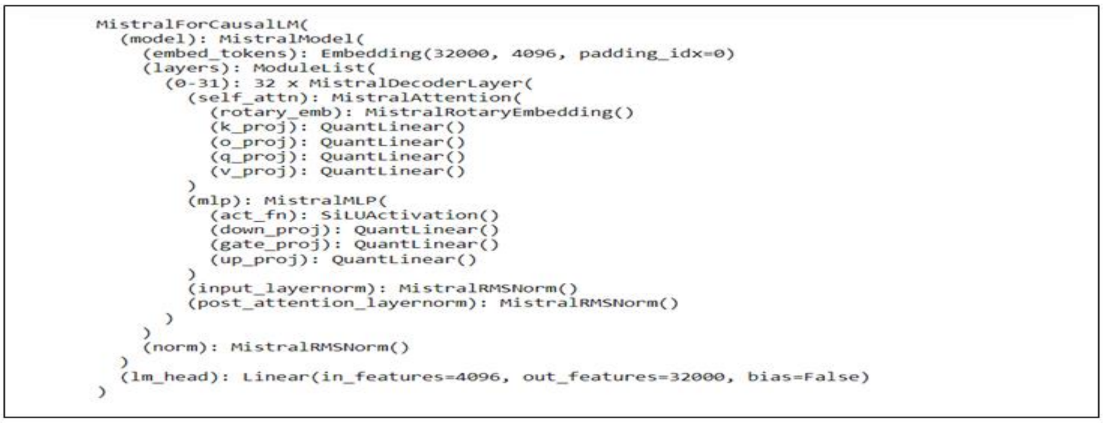




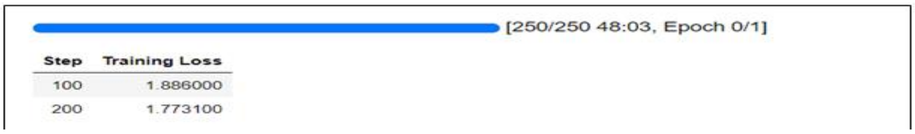
3. News Summarization:

We compared Mistral 7B to the Llama 2 family, and re-run all model evaluations ourselves for fair comparison





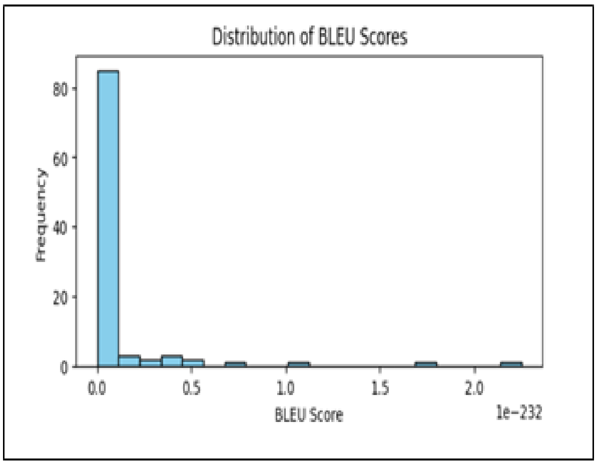
```
MistralForCausalLM(
  (model): MistralModel(
    (embed_tokens): Embedding(32000, 4096, padding_idx=0)
    (layers): ModuleList(
      (0-31): 32 x MistralDecoderLayer(
        (self_attn): MistralAttention(
          (rotary_emb): MistralRotaryEmbedding()
          (k_proj): QuantLinear()
          (o_proj): QuantLinear()
          (q_proj): QuantLinear()
          (v_proj): QuantLinear()
        )
        (mlp): MistralMLP(
          (act_fn): SiLUActivation()
          (down_proj): QuantLinear()
          (gate_proj): QuantLinear()
          (up_proj): QuantLinear()
        )
        (input_layernorm): MistralRMSNorm()
        (post_attention_layernorm): MistralRMSNorm()
      )
    )
    (norm): MistralRMSNorm()
  )
  (lm_head): Linear(in_features=4096, out_features=32000, bias=False)
)
```

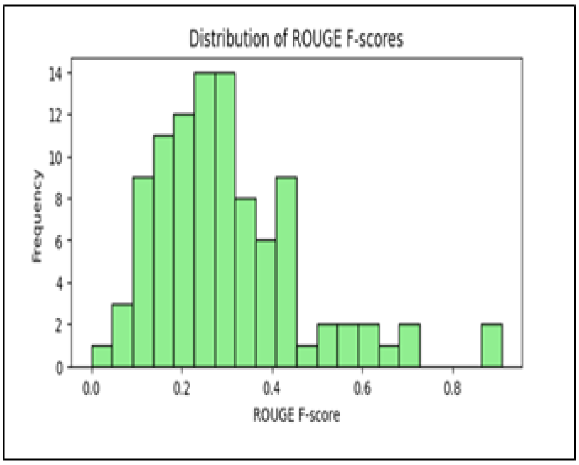


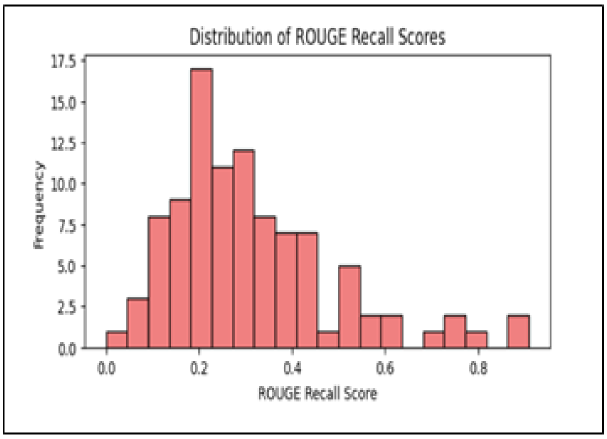
```
[250/250 48:03, Epoch 0/1]

Step   Training Loss
100    1.886000
200    1.773100
```

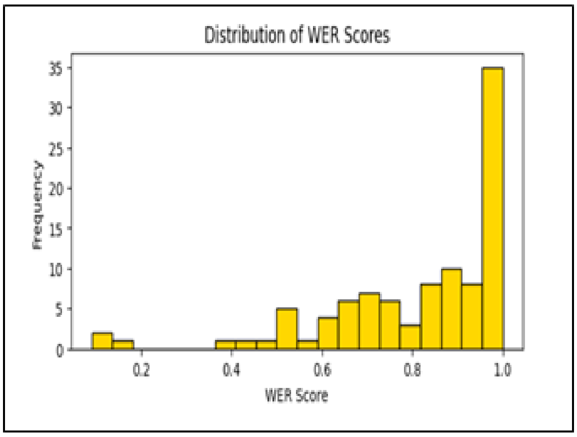
4. News Translation English To Hindi:

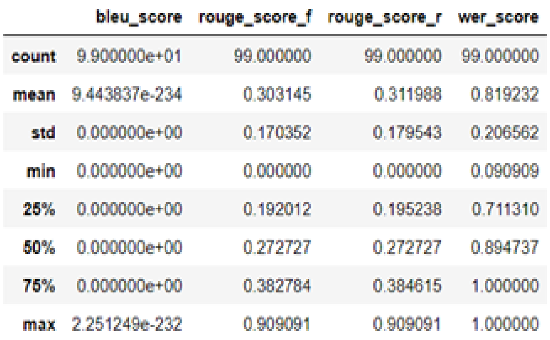
- BLEU Score: 9.44%
- ROUGE F-Score: 30%
- ROUGE R-Score: 31%
- WER Score: 81%







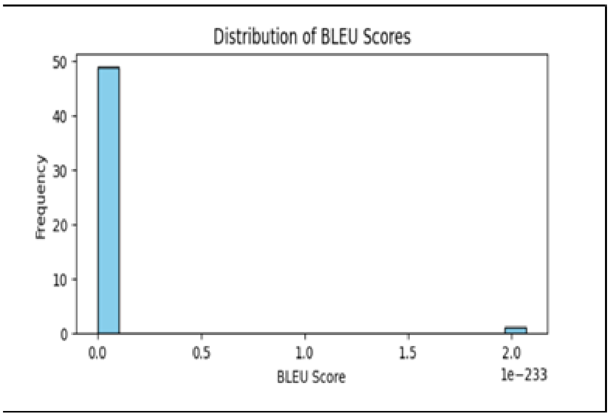


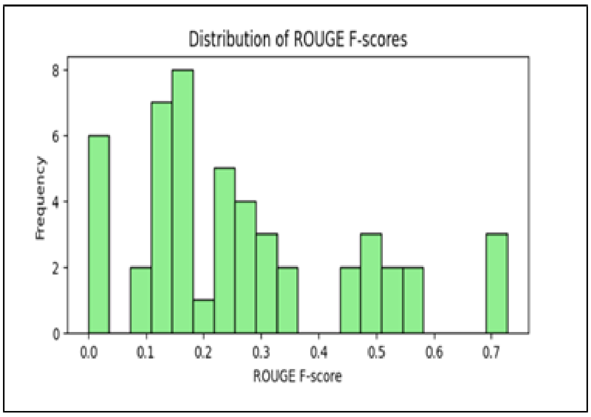


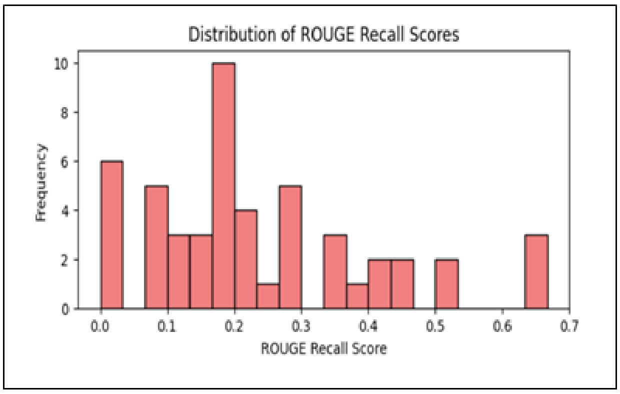
	bleu_score	rouge_score_f	rouge_score_r	wer_score
count	9.900000e+01	99.000000	99.000000	99.000000
mean	9.443837e-234	0.303145	0.311988	0.819232
std	0.000000e+00	0.170352	0.179543	0.206562
min	0.000000e+00	0.000000	0.000000	0.090909
25%	0.000000e+00	0.192012	0.195238	0.711310
50%	0.000000e+00	0.272727	0.272727	0.894737
75%	0.000000e+00	0.382784	0.384615	1.000000
max	2.251249e-232	0.909091	0.909091	1.000000

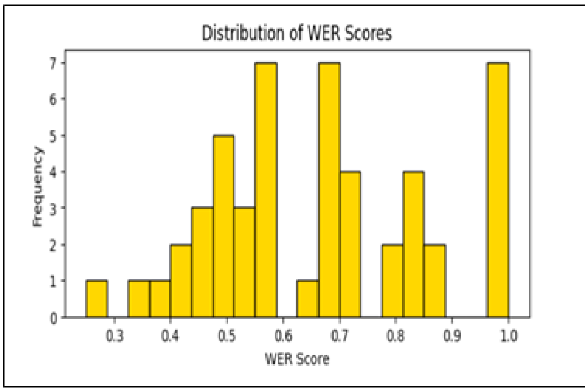
5. News Translation English To Marathi:

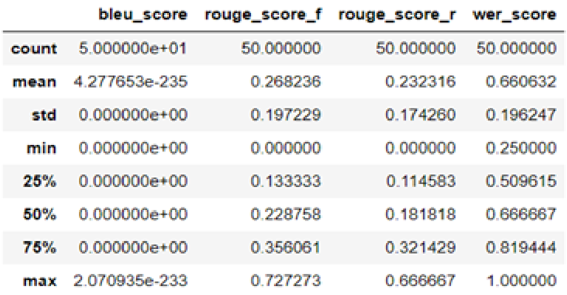
- BLEU Score: 4.27%
- ROUGE F-Score: 26%
- ROUGE R-Score: 23%
- WER Score: 66%











	bleu_score	rouge_score_f	rouge_score_r	wer_score
count	5.000000e+01	50.000000	50.000000	50.000000
mean	4.277653e-235	0.268236	0.232316	0.660632
std	0.000000e+00	0.197229	0.174260	0.196247
min	0.000000e+00	0.000000	0.000000	0.250000
25%	0.000000e+00	0.133333	0.114583	0.509615
50%	0.000000e+00	0.228758	0.181818	0.666667
75%	0.000000e+00	0.356061	0.321429	0.819444
max	2.070935e-233	0.727273	0.666667	1.000000

VIII. CONCLUSION

News Nexa: Beyond Headlines uses a variety of deep learning models for Natural Language Processing (NLP) activities to address the problems of information overload and disinformation in the digital age. Users' access to and interaction with news content could be revolutionized by this project, which has the potential to empower them.

Principal Accomplishments: Deep learning models (RNNs, LSTMs, GRUs, BERT, T5) were created and compared for the classification of fake news, news category classification, and news summarization. Transformers and GPT-2 models were used to create succinct and educational news story summaries. investigated the Mistral-7B LLM's potential for additional improvements in summarization. Transformers and GPT-2 models were used to establish an English to Hindi news translation system. looked into the viability of using open-source translation libraries to translate from English to Marathi.

ACKNOWLEDGEMENTS

This is our chance to express our gratitude to our project guide Prof. Ms. Gaikwad S. G., as well as our honorary principal, Prof. Dr. Mukane S. M., for their invaluable advice and for providing all the resources that were required to finish this project report,. We also appreciate all of the staff of the SVPM College of Engineering, Malegaon (Bk), Department of Information Technology, for their invaluable time, assistance, remarks, recommendations, and persuasion. We express our gratitude to the institute for furnishing necessary amenities, internet connectivity, and significant literature.

IX. REFERENCES

[1] Shahbaz, M., et al. (2019). Fake News Detection Based on Machine Learning Algorithms. https://ieeexplore.ieee.org/document/9378748/

[2] Zhang, Y., et al. (2015). Character-Level Convolutional Networks for Text Classification. https://arxiv.org/abs/1509.01626 (This explores character-level CNNs, a possible approach for text classification)

[3] Hassan, A., et al. (2017). Achievable Accuracy and Fairness in Text Classification. https://arxiv.org/pdf/2309.13761

[4] Liu, Y., et al. (2019). GPT-2 Summarization with Pointer-Generator Network. https://arxiv.org/pdf/1905.01975 (This explores GPT-2 for summarization with a specific technique)

[5] Nallapureddy, K., & Xiang, B. (2017). Abstractive Summarization with Attentive LSTMs. https://aclanthology.org/N16-1012

[6] Rush, A. M., et al. (2015). Neural Attention for Sequence-to-Sequence Learning.

[7] https://arxiv.org/abs/1409.3215 (This one explores attention mechanisms, crucial for summarization models)

[8] Devlin, J., et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics.

[9] Vaswani, A., et al. (2017). Attention Is All You Need. [https://arxiv.org/abs/1706.03762] (This is the foundational paper on Transformers, the architecture used for GPT-2 and many translation models)

[10] Albert Q. Jiang.,(2023). Mistral 7B [https://arxiv.org/abs/2310.06825]