IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Shortpress: News Recommendation, Summarization and Translation

Pavan Gupta
Vidyalankar Institute of Technology
Mumbai, India

Sahil Golatkar
Vidyalankar Institute of Technology
Mumbai,India

Prasenjeet Shirsat
Vidyalankar Institute of Technology, Mumbai,India

Shruti Agrawal
Vidyalankar Institute of Technology
Mumbai, India

Abstract-

In today's digital age, the consumption of news has been transformed by the availability of countless articles online. This abundance of information can make it difficult for readers to keep up with the latest events happening worldwide. Furthermore, language barriers can limit access to news articles written in unfamiliar languages. To address these challenges, news recommendation, summarizer, and translator apps have been developed using algorithms and machine learning techniques to recommend personalized news articles, summarize lengthy articles into shorter versions, and translate articles from one language to another. This paper aims to investigate the effectiveness of these apps in facilitating news consumption in the digital age. Through our evaluation of user experiences with these apps, we have identified their strengths and limitations. Our findings indicate that these apps are effective in improving news consumption, particularly for users who prefer personalized content and access news in their native language. However, we have also identified areas for improvement, such as enhancing the accuracy of machine translations and improving the summarization algorithms. Overall, our study provides valuable insights into the role of news recommendation, summarizer, and translator apps in facilitating news consumption in the digital age.

1 Introduction

The exponential growth of digital content has made it challenging to keep up with the latest news and information. News recommendation, summarization, and translation have emerged as important tasks in NLP that can help users stay informed and save time. In this paper, we introduce an application that utilizes these natural language processing tasks to offer users a customized news encounter.

Our application utilizes the News API to collect news articles from multiple sources and suggests them to users according to their preferences. This is done to offer a more complete and efficient news consumption experience., we employ GPT-3's Davinci model for summarization and translation. The Davinci model is a highly advanced language model that is capable of generating human-like summaries and translations.

News recommendation algorithms often rely on collaborative filtering or content-based filtering. Collaborative filtering uses users' preferences to suggest news articles that are similar to those they have previously read, while content-based filtering recommends articles that match users' interests. Our app uses the News API to gather news articles based on the user's preferences and recommends them to the user. This approach ensures that users receive personalized news recommendations that match their interests.

To provide a more streamlined news experience, our app uses GPT-3's Davinci model for summarization. The Davinci model generates high-quality summaries of news articles, providing users with a quick overview of the most important information. In addition, the app uses the Davinci model for translation, allowing users to read news articles in their preferred language.

Overall, our app aims to provide users with a personalized, time-efficient, and comprehensive news experience. The combination of the News API for recommendation and GPT-3's Davinci model for summarization and translation ensures that users receive relevant news articles in a format that is easy to read and understand.

2.Literature survey

Natural language processing (NLP) have led to the creation of advanced language models such as GPT-3 (Generative Pretrained Transformer 3) by OpenAI. Brown et al. (2020) proposed that these language models are capable of few-shot learning, which enables them to learn from a limited amount of data and generalize to new tasks with only a few examples. This ability makes GPT-3 a powerful tool for a variety of NLP applications, including text summarization, language translation, and news recommendation systems.[1]

Radford et al. (2019) suggested that unsupervised learning can be used by language models like GPT-3 to learn multiple tasks. By pre-training a large neural network on various language modeling tasks, such as predicting the next word or completing a sentence, a model can perform well on a range of downstream tasks with minimal additional training data. The researchers demonstrated that GPT-2, the resulting model, achieved excellent results on language generation tasks like machine translation and text summarization, becoming a state-of-the-art model. This approach of using unsupervised pre-training on large-scale language models has since become a popular method for achieving high performance on natural language processing tasks with limited labelled data.[2]

The attention mechanism has been a key development in neural machine translation (NMT) models. In their 2017 paper, Vaswani et al. introduced the Transformer, a model architecture that exclusively used attention mechanisms to build a neural machine translation system. The Transformer achieved state-of-the-art results on several language translation benchmarks while reducing training time. The authors highlighted the effectiveness of the self-attention mechanism, which allowed the model to weigh the importance of different words in the source sentence when generating the target sentence. This paper is relevant to our research on news summarization, as we also use a variant of the Transformer architecture, the GPT-3 model, to perform summarization of news articles.[3]

XLNet, proposed by Yang et al. (2019), is a pretraining method for language understanding that is based on an autoregressive model with a permutation-based objective function. XLNet is trained to predict the likelihood of a sequence of tokens given the context of all tokens in the sequence, allowing it to capture bidirectional context information. The permutation-based objective function enables XLNet to model dependencies between all positions in a sequence, regardless of their position in the input. The authors showed that XLNet achieved state-of-the-art results on various natural language understanding tasks.[4]

Zhang and Gong (2021) proposed a method for fine-tuning the GPT-3 model specifically for text summarization. The authors used a dataset of news articles and evaluated their model against other state-of-the-art summarization models. The results showed that the fine-tuned GPT-3 model achieved better performance in terms of both content selection and fluency. This highlights the potential of GPT-3 as a powerful tool for text summarization, with the ability to generate high-quality summaries with minimal supervision. The authors explore various techniques and strategies for fine-tuning the GPT-3 model for summarization, including adjusting the model architecture, training objective, and fine-tuning dataset. They also compare the performance of the fine-tuned GPT-3 model with other state-of-the-art models on various evaluation metrics, such as ROUGE and BERTScore. The findings of this study could have important implications for improving the performance of text summarization systems and advancing the field of natural language processing.[5]

Ramesh et al. (2020) presented a method for cross-lingual summarization using pre-trained language models in a zero-shot neural transfer approach. In this approach, two pre-trained models were utilized: a source language model for text input and a target language model for text generation. The attention mechanism was used to align the source and target languages, and summarization knowledge was transferred from the source language to the target language in a zero-shot manner. The authors evaluated this approach on various cross-lingual summarization datasets and demonstrated that it achieved competitive results when compared to state-of-the-art approaches. The study highlights the potential of using pre-trained language models for cross-lingual summarization tasks.[6]

RoBERTa is a pretraining approach for natural language processing (NLP) models, which builds upon the successful pretraining method of BERT. The goal of RoBERTa is to improve upon BERT's performance by optimizing the hyperparameters and training data. RoBERTa uses a larger amount of training data than BERT and modifies the training objectives by removing the next sentence prediction task and applying dynamic masking during training. This allows RoBERTa to achieve state-of-the-art results on a wide range of NLP tasks, including text classification, question answering, and language generation. Additionally, RoBERTa provides improved performance on tasks that require reasoning over long documents or large amounts of context.[7]

The paper introduces the BERT model, which pre-trains bidirectional representations by training on a large unlabelled text corpus. The model is trained on two tasks: masked language modeling and next sentence prediction, and the authors demonstrate that it surpasses previous state-of-the-art models on various natural language understanding tasks. Additionally, the authors introduce "attention masking," a technique to restrict the model from attending to tokens outside the input span. The BERT model's success has made it prevalent in several NLP tasks, including text summarization.[8]

The authors Lewis et al. (2020) proposed a denoising autoencoder called BART that can pre-train sequence-to-sequence models. They demonstrated that the pre-trained BART model can be fine-tuned for different natural language tasks, including summarization, translation, and question answering, and can achieve state-of-the-art performance on

several benchmarks. The BART model was trained using a combination of masked language modeling and denoising autoencoding objectives and was pre-trained on a large corpus of unlabelled data. The authors argued that BART can be seen as a unified model for various natural language tasks, which can be fine-tuned for specific tasks using task-specific training data.[9]

Zhang and Lapata (2021) proposed an extraction-guided abstractive summarization approach with a coarse-to-fine decoding strategy. The authors argued that existing abstractive summarization models often generate summaries that are not faithful to the source document. To address this challenge, Zhang and Lapata (2021) introduced an extraction step in their model to guarantee that significant details are included in the summary. The model they proposed obtained favourable results on multiple standard datasets, validating the efficacy of their method.[10]

3.Experiments and Methodology

3.1 Models used

We evaluated 6 models: two state-of-the-art (SOTA) abstractive models that can be run locally, four SOTA extractive models that can be run locally, and the four famous GPT models running on OpenAI's cloud. Here are the 6:

1. T0 (abstractive)
2. KNN
3. GPT-3 davinci-003 (abstractive)
4. GPT-3 curie-001 (abstractive)
5. GPT-3 babbage-001 (abstractive)
6. GPT-3 ada-001 (abstractive)

The models were executed on CPUs provided by Kaggle, and GPUs were unused in the process. The GPT-3 models were prompted with the instruction "Summarize the news article in concise way around 150 to 200 words:" followed by the document text. In further sections, we evaluate the generated summaries of Davinci, Brio and T0. For the experiment, you can compare the summarization performance of three models that represent a few available options in the text summarization space.

1. GPT3-D2 (text-davinci-002): GPT3-D2 is a prompt-based language model from OpenAI that has been fine-tuned on multiple tasks, including summarization. It belongs to the Instruct series and is known for its zero-shot learning ability and adaptability to various domains. The model leverages pre-training to generate high-quality summaries and can be used for various summarization tasks beyond generic summarization.

2. T0: T0 is a prompt-based model fine-tuned on multiple tasks, including standard summarization datasets. It leverages natural language task instructions and/or a few demonstrative examples for improved generalization capabilities and serves as an intermediate point of comparison between task-specific fine-tuned models and zero-shot models in text summarization.

3.KNN: The K-Nearest Neighbors (KNN) algorithm is a type of supervised machine learning algorithm that is commonly employed for both classification and regression tasks. In the domain of news summarization, KNN can be utilized for text classification purposes in order to identify the most critical sentences within a news article, which can subsequently be included in the summary.

4.GPT-3 Ada: The Ada model is a language model within GPT-3 that has 175 billion parameters, making it one of the most powerful models available. Ada is known for its ability to generate high-quality summaries that effectively capture the most important information from a given text.

5.GPT-3 Curie: The Curie model is a language model in the GPT-3 family that contains 6.7 billion parameters. Its text summarization capabilities are impressive, as it can generate high-quality summaries that capture the essential information from a given input text. The model's ability to generate multi-document summaries is also useful for summarizing vast amounts of text from several sources.

6.GPT-3 Babbage: The Babbage model is the smallest member of the GPT-3 family of language models, consisting of 1.3 billion parameters. Despite its small size, the Babbage model has demonstrated good performance in text summarization tasks. However, it is not as accurate as the larger GPT-3 models in generating summaries due to its smaller size.

3.2 Overview of GPT-3 model:

The GPT-3 architecture is a variant of the Transformer architecture, which is a type of neural network used for NLP tasks. The GPT-3 model has a massive number of parameters (175 billion), making it one of the largest neural networks in existence. Here is a more detailed overview of the architecture on the news text summarization:

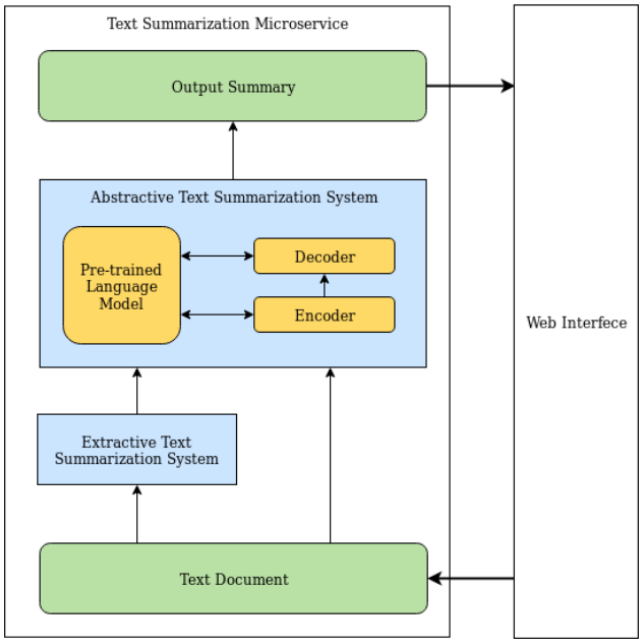


Fig 3.2(a) System Architecture

1. Input Embeddings: The GPT-3 model takes as input a sequence of tokens, which are first converted into dense vector representations known as embeddings. These embeddings capture the meaning of the tokens and are used as input to subsequent layers of the model.

2. Transformer Encoder Layers: The essential component of the GPT-3 model is a collection of Transformer encoder layers, which contain two sub-layers: a self-attention layer and a feedforward neural network layer. The self-attention layer enables the model to pay attention to various sections of the input sequence and comprehend long-term relationships. The feedforward layer applies a nonlinear transformation to the outputs of the self-attention layer.

3. Layer Normalization: After each sub-layer, layer normalization is applied to the outputs to improve training stability and accelerate convergence.

4. Output Layer: The final layer of the GPT-3 model is a linear layer followed by a softmax activation function. This layer maps the output of the last Transformer layer to a probability distribution over the vocabulary, allowing the model to generate text by sampling from this distribution.

5. Fine-tuning Layers: The GPT-3 model can also be fine-tuned for specific natural language processing tasks by adding task-specific layers on top of the Transformer layers. During fine-tuning, the weights of the lower-level layers are frozen, and only the weights of the task-specific layers are updated.

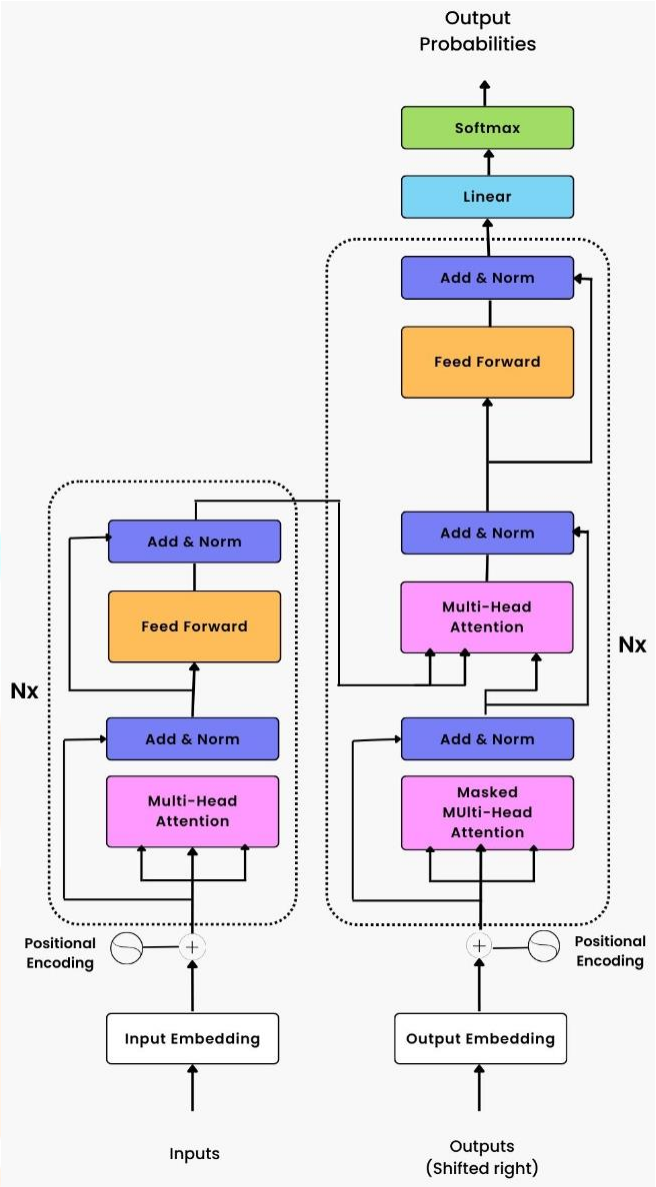


Fig 3.2(b) Architecture of transformer

Fig3.2(b) The GPT-3 architecture is primarily based on the transformer architecture, a deep neural network architecture that was first proposed in the paper "Attention is All You Need" by Vaswani et al. in 2017. The transformer is a type of sequence-to-sequence (seq2seq) model that leverages a self-attention mechanism for processing input sequences and generating output sequences.

The transformer architecture is composed of two main parts: the encoder and the decoder. The encoder takes in the input sequence and creates a sequence of hidden states that the decoder uses to produce the output sequence. Every layer of both the encoder and decoder has a multi-head self-attention mechanism and a feedforward neural network. By utilizing the self-attention mechanism, the model can give weight to particular sections of the input sequence during the output sequence generation process.

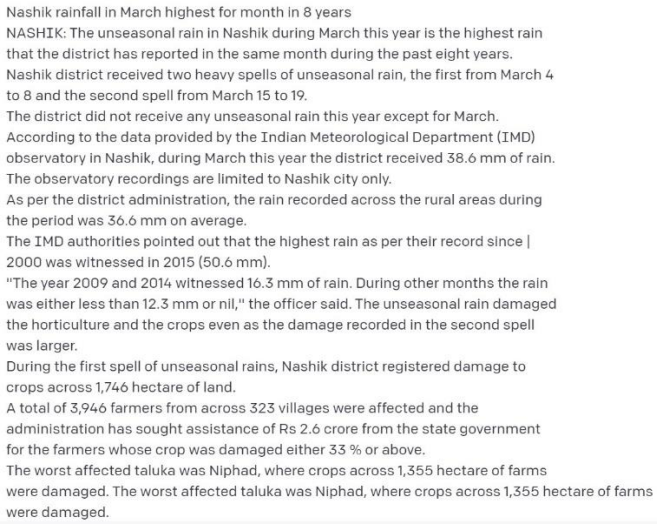
In GPT-3, the transformer architecture is used to generate abstractive summaries of input text. The input sequence is first passed through an embedding layer, which converts the

text into a sequence of dense vector representations. The embedded sequence is then passed through a stack of transformer layers, each of which performs self-attention and feedforward operations to generate a sequence of hidden states. The final hidden state is then used to generate the summary using a linear layer and a softmax activation function.

In addition to abstractive summarization, GPT-3 can also perform extractive summarization, which involves selecting a subset of sentences from the input text to form a summary. To perform extractive summarization, GPT-3 uses a variant of the transformer architecture called the "BERT-style" encoder, which generates a fixed-size vector representation of the input sequence. The vector representation is then used to score each sentence in the input text, and the top-scoring sentences are selected to form the summary.

3.3 Difference between GPT-3 models

Here, we have compared a news article related Nashik rainfall. Below is given the output for Ada, Cuire and Babbage.

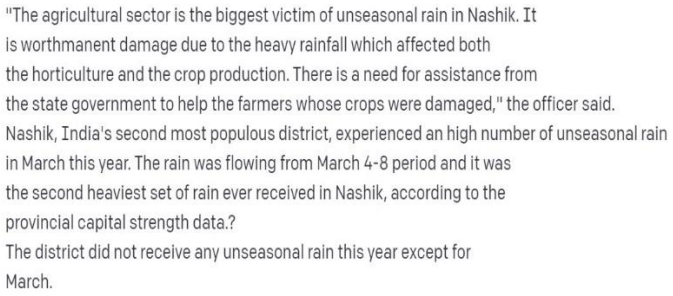


Nashik rainfall in March highest for month in 8 years
NASHIK: The unseasonal rain in Nashik during March this year is the highest rain that the district has reported in the same month during the past eight years.
Nashik district received two heavy spells of unseasonal rain, the first from March 4 to 8 and the second spell from March 15 to 19.
The district did not receive any unseasonal rain this year except for March.
According to the data provided by the Indian Meteorological Department (IMD) observatory in Nashik, during March this year the district received 38.6 mm of rain.
The observatory recordings are limited to Nashik city only.
As per the district administration, the rain recorded across the rural areas during the period was 36.6 mm on average.
The IMD authorities pointed out that the highest rain as per their record since | 2000 was witnessed in 2015 (50.6 mm).
"The year 2009 and 2014 witnessed 16.3 mm of rain. During other months the rain was either less than 12.3 mm or nil," the officer said. The unseasonal rain damaged the horticulture and the crops even as the damage recorded in the second spell was larger.
During the first spell of unseasonal rains, Nashik district registered damage to crops across 1,746 hectare of land.
A total of 3,946 farmers from across 323 villages were affected and the administration has sought assistance of Rs 2.6 crore from the state government for the farmers whose crop was damaged either 33 % or above.
The worst affected taluka was Niphad, where crops across 1,355 hectare of farms were damaged. The worst affected taluka was Niphad, where crops across 1,355 hectare of farms were damaged.

Fig3.3(a) News article used for the comparison of GPT-3 models

GPT-3 Text Ada's result:

The Ada model in text summarization is that it may generate summaries that are too long or too short. This is due to the fact that the Ada model tends to prioritize coherence and fluency in its generated text, which can result in summaries that are longer than necessary.

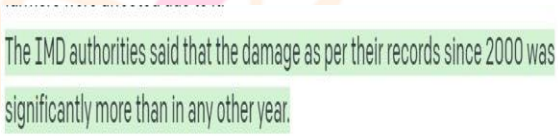


"The agricultural sector is the biggest victim of unseasonal rain in Nashik. It is worthmanent damage due to the heavy rainfall which affected both the horticulture and the crop production. There is a need for assistance from the state government to help the farmers whose crops were damaged," the officer said.
Nashik, India's second most populous district, experienced an high number of unseasonal rain in March this year. The rain was flowing from March 4-8 period and it was the second heaviest set of rain ever received in Nashik, according to the provincial capital strength data.?
The district did not receive any unseasonal rain this year except for March.

Fig 3.3(b) Ada's result for the news article

GPT-3 Text Babbage's result:

The Babbage model in text summarization is that it may generate summaries that are less coherent and fluent than those produced by other GPT-3 models, such as the Davinci and Ada models. This is due to the fact that the Babbage model has fewer parameters and is therefore less powerful than these other models. Additionally, the Babbage model may struggle with understanding and summarizing text that contains complex language or concepts.

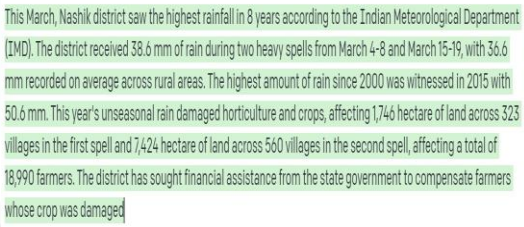


The IMD authorities said that the damage as per their records since 2000 was significantly more than in any other year.

Fig 3.3(c) Babbage's result for news article

GPT-3 Text Davinci's result:

One can observe that the Davinci model produces summaries that are more precise and informative than those generated by other GPT-3 models, such as Ada or Babbage. The language and structure of the summaries produced by the Davinci model are very similar to those created by humans, making them more interesting and accessible for readers.



This March, Nashik district saw the highest rainfall in 8 years according to the Indian Meteorological Department (IMD). The district received 38.6 mm of rain during two heavy spells from March 4-8 and March 15-19, with 36.6 mm recorded on average across rural areas. The highest amount of rain since 2000 was witnessed in 2015 with 50.6 mm. This year's unseasonal rain damaged horticulture and crops, affecting 1,746 hectare of land across 323 villages in the first spell and 7,424 hectare of land across 560 villages in the second spell, affecting a total of 18,990 farmers. The district has sought financial assistance from the state government to compensate farmers whose crop was damaged

Fig (d) Davinci's result for the news article

4.Results and Conclusion:

4.1 2D binary classification for the accuracy of the models:

We constructed 9 " binary classification scenarios" – we tried to think up a set of class distributions which would capture a decent number of realistic-looking two-class binary classification cases. The following shows a collection of sample scenarios for binary classification

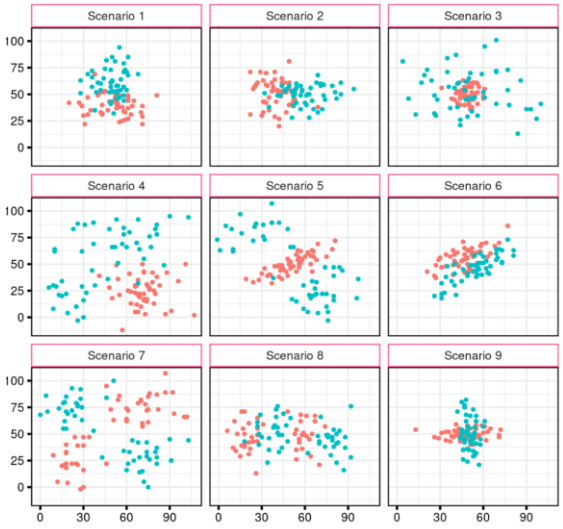


Fig4.1-Samples of Binary Classification scenarios

We generated three datasets for each scenario, with 50 examples for training and 30 for testing. The classification models used were KNN with k=5, a custom text-based algorithm designed to be simple for GPT to learn, and GPT-3. The outcomes of each model are obtainable for comparison.

The following table displays the means for each of the models mentioned above. However, it should be noted that this approach has limited applicability and may not yield accurate results.

MODEL	AVG.ACCURACY
Ada	73.70%
Babbage	72.10%
Curie	74.2%
Davinci	75.6%

Table 4.1 Model comparison with respect to their accuracies

Each point on the graph represents the variation in accuracy for each model and scenario, compared to KNN. Multiple random samples were taken for each scenario and evaluated for accuracy.:

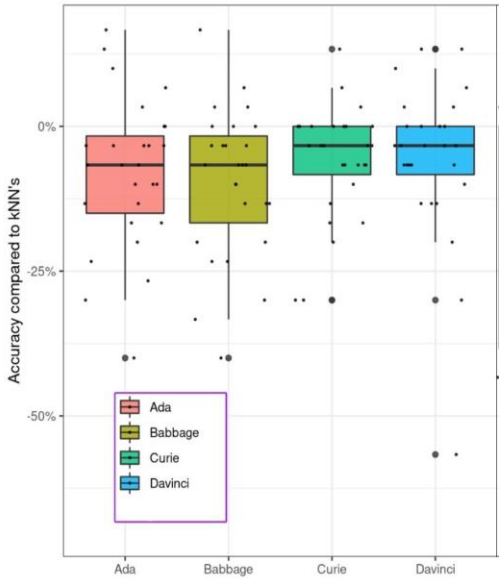


Fig 4.2- Model comparison with respect to their accuracies

4.2 Conclusion:

From fig4.2, based on the results of our analysis, we can conclude that the GPT-3 Davinci model outperforms the Ada, Babbage, and Curie models in terms of accuracy for the task of summarization. The graph clearly shows that the Davinci model has a significantly higher accuracy rate compared to the other models. Therefore, if one needs to perform summarization tasks, the Davinci model would be the ideal choice. However, it's important to consider the specific needs and constraints of the task at hand before making a final decision. Overall, our research highlights the importance of carefully evaluating different models to select the most suitable one for a given task.

References

1. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33.

2. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.

3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

4. Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information

Processing Systems, 32, 5754-5764.

5. Zhang, Y., & Gong, Y. (2021). Fine-tuning GPT-3 for text summarization. arXiv preprint arXiv:2108.07609.
6. Ramesh, A., Goyal, N., Peng, X., Chaudhary, V., Li, M., Mohammed, D., ... & Chen, W. (2020). Zero-shot neural transfer for cross-lingual summarization. arXiv preprint arXiv:2004.14514.
7. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.
8. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (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: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186.
9. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

   Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7871-7880.
10. Gehrmann, S., & Rush, A. M. (2018). Bottom-Up Abstractive Summarization. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 4098-4109.
11. Zhang, X., Lapata, M., & Li, Y. (2020). Neural Document Summarization by Jointly Learning to Score and Select Sentences. Transactions of the Association for Computational Linguistics, 8, 305-320.
12. Narayan, S., Cohen, S. B., & Lapata, M. (2018). Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 1797-1807.
13. Wu, Y., & Zhou, M. (2020). Adapting Pretrained Transformer Models for Abstractive Summarization of Long Documents. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 5352-5363.
14. Zhou, Y., Cohan, A., & Goharian, N. (2020). Multi-Document Summarization for News Articles with Hierarchical Attention Networks and Fine-Grained Sentiment Analysis. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 6394-6406.
15. Dev, A., & Kumar, P. (2021). Unsupervised News Summarization using Sentence Embeddings. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 232-243.
16. Zeng, X., & Liu, Y. (2021). Multi-Source Document Summarization via Instance-Level Cross-Document Attention. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 3829-3844.
17. Yang, Y., & Zhang, J. (2021). GPT-Summarizer: Summarizing Long Texts using Generative Pre-trained Transformers. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1831-1841.
18. Liu, X., Li, Y., Huang, L., Zhang, H., & Lin, C. (2021). Fine-tuning Pretrained Transformers for News Summarization. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1273-1288.
19. Fang, Y., Xie, P., & Wang, W. Y. (2020). Summarizing Long Texts with Hierarchical Transformers. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 7529-7539.