AI VS. HUMAN: EXPLORING THE LIMITS OF MACHINE INTELLIGENCE

MOTIVATION

Human-generated question and answer (Q&A) platforms have been popular for a long time now, providing a wealth of information and knowledge on a wide variety of topics. To provide answers to users' questions, these platforms rely on user-generated content. They are an invaluable resource for research, education, and personal growth due to the diversity of perspectives and expertise they provide. On the other hand, in today's digital age, generative AI models have emerged as a rapidly growing and significant means of human-computer interaction. These platforms leverage machine learning algorithms to generate responses to user queries and provide a human-like conversational experience. Chatbots, virtual assistants, and customer service bots are examples of such algorithm-generated platforms that are becoming more common.

As Artificial Intelligence becomes increasingly sophisticated, there is a growing concern that it might eventually replace humans in many domains, including the creation of written content such as articles, blogs, and answers to frequently asked questions on websites like Quora. Our research specifically targets to differentiate AI-generated responses from ChatGPT and human-generated responses on Quora. Comparison of ChatGPT answers and human answers is done which is further used for the analysis to determine the areas in which AI is most likely to produce human-like responses.

This research is essential for industry practitioners and researchers who are concerned with the impact of AI on various fields, as well as for the advancement of knowledge in the field of AI and ML. By understanding the capabilities and limitations of AI in generating written content, we can better prepare for the future of industries and leverage the potential of AI for enhanced productivity and innovation.

RESEARCH QUESTIONS

❖ Can a binary classification model accurately distinguish between answers generated by ChatGPT and those generated by humans?

❖ What features of the answers are most useful in predicting whether an answer was generated by ChatGPT or by a human?

❖ Are ChatGPT-generated answers comparable to human-generated answers on Quora?

❖ In what categories of questions is ChatGPT most effective at generating similar answers to humans?

BACKGROUND AND RELATED WORK

In recent years, there has been growing interest in differentiating between AI-generated and human-generated answers. Several studies have been conducted on this topic, with different approaches and methodologies. (Gehrmann, Rush and Strobelt, 2019) have developed GLTR, a tool that automatically detects and visualizes the properties of text that correlate with the likelihood of being synthetic. The study found that the use of GLTR enabled untrained humans to more accurately detect synthetic text from 54% to 72%. In another study, (Mitrović, Andreoletti and Ayoub, 2023) focused on comparing human-generated short online reviews with the ones generated by ChatGPT using a ML model. The study has shown that the ML model overpowers traditional methods such as perplexity, by achieving an accuracy of 79%.

While these studies have made significant progress in the detection of AI-generated text, there is still a need for further research to improve the accuracy of classification models and to explore the potential applications of such models in various fields. Therefore, our present study aims to contribute to this growing body of literature by developing and evaluating a model for comparing AI-generated answers with human-generated answers. This research is particularly relevant in the current age of rapidly evolving AI technology, where the ability to accurately differentiate between AI-generated and human-generated text is becoming increasingly important.

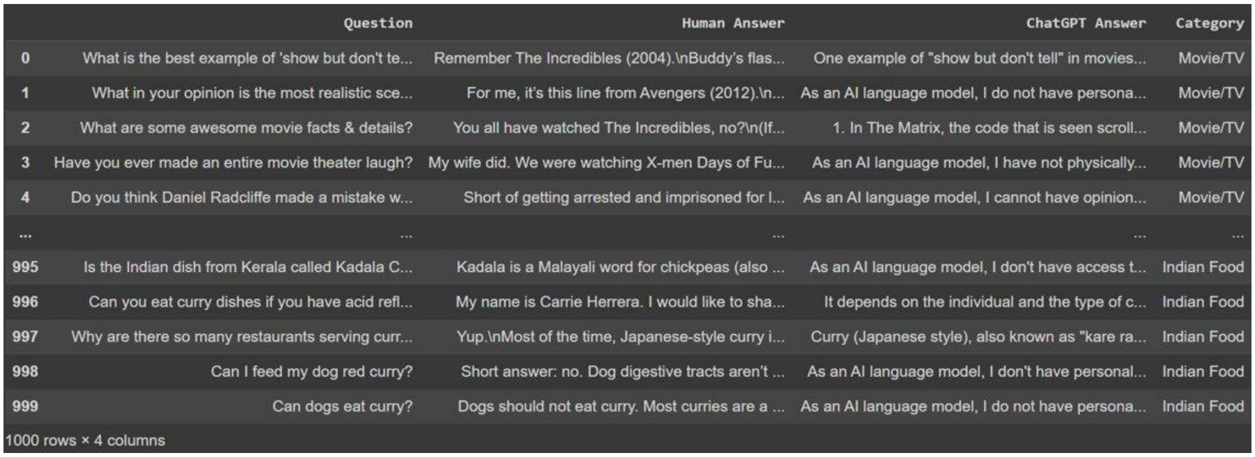
METHODOLOGY

DATASET COLLECTION

The dataset is collected through a two-step process:

1) Scraping questions and human generated answers from Quora
   A list of web pages are provided, from where the data is to be scraped. To ensure collection of maximum data, the web page is scrolled until the end of the page. The answers are expanded wherever necessary, to ensure capturing of complete responses. To avoid duplicate data, questions that have already been answered, and those with insufficient or incomplete responses are skipped. Additionally, only the questions with answers longer than a given threshold are included. Finally, the collected data is stored in an excel file and a category column is added to help organize the data. This process is repeated until a sufficient amount of data is collected or all the web pages in the list are scraped, whichever comes first.
2) Collecting answers to the scraped questions from ChatGPT.
   The scraped questions are sequentially fed into ChatGPT using the OpenAI API, and the resulting answers are appended to a column in the existing data.

The collected dataset includes a total of 1000 questions and their corresponding answers, with 250 samples collected from each of the four domains: Movie/TV, Science/Tech, Philosophy, and Indian Food.



	Question	Human Answer	ChatGPT Answer	Category
0	What is the best example of 'show but don't te...	Remember The Incredibles (2004).\nBuddy's flas...	One example of "show but don't tell" in movies...	Movie/TV
1	What in your opinion is the most realistic sce...	For me, it's this line from Avengers (2012).\n...	As an AI language model, I do not have persona...	Movie/TV
2	What are some awesome movie facts & details?	You all have watched The Incredibles, no?\n(If...	1. In The Matrix, the code that is seen scroll...	Movie/TV
3	Have you ever made an entire movie theater laugh?	My wife did. We were watching X-men Days of Fu...	As an AI language model, I have not physically...	Movie/TV
4	Do you think Daniel Radcliffe made a mistake w...	Short of getting arrested and imprisoned for l...	As an AI language model, I cannot have opinion...	Movie/TV
...	...	...	...	...
995	Is the Indian dish from Kerala called Kadala C...	Kadala is a Malayali word for chickpeas (also ...	As an AI language model, I don't have access t...	Indian Food
996	Can you eat curry dishes if you have acid refl...	My name is Carrie Herrera. I would like to sha...	It depends on the individual and the type of c...	Indian Food
997	Why are there so many restaurants serving curr...	Yup.\nMost of the time, Japanese-style curry i...	Curry (Japanese style), also known as "kare ra...	Indian Food
998	Can I feed my dog red curry?	Short answer: no. Dog digestive tracts aren't ...	As an AI language model, I don't have personal...	Indian Food
999	Can dogs eat curry?	Dogs should not eat curry. Most curries are a ...	As an AI language model, I do not have persona...	Indian Food

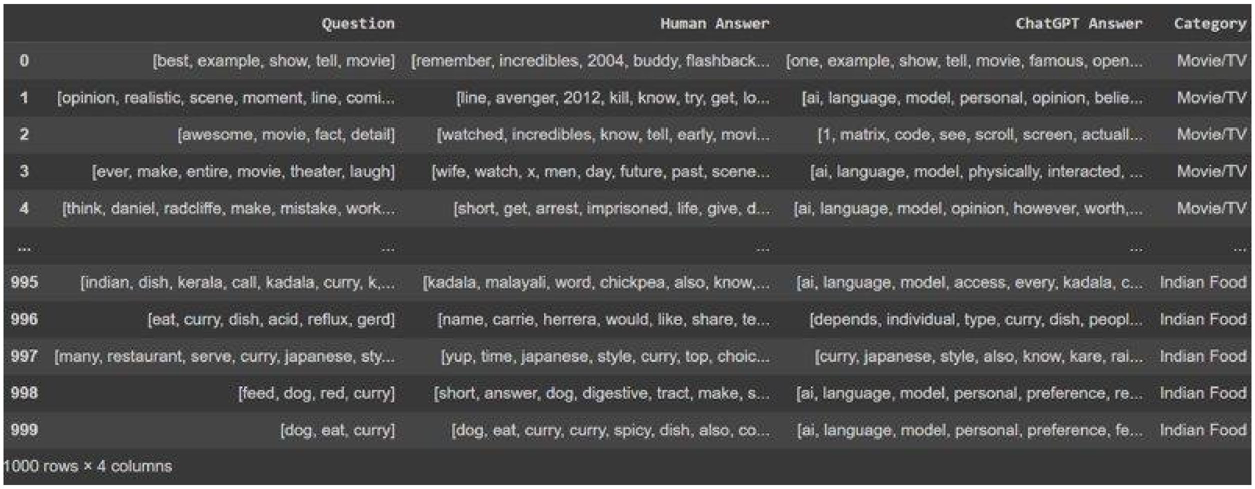
1000 rows × 4 columns

DATA PREPROCESSING

Data preprocessing is a crucial step in the data analysis pipeline, as it helps to transform the raw text data into an easier format for analysis. This step involves extracting relevant information from the question-answer pairs utilizing various natural language processing (NLP) techniques that are commonly used for text pre-processing.

The scraped data is loaded and text pre-processing techniques are applied to prepare it for further analysis. To achieve this, a function is defined where the pre-processing techniques are applied based on the arguments passed. This approach allows for flexibility in the pre-processing steps and enables easy experimentation with different techniques. The function can be called with different arguments to perform various pre-processing steps such as lowercasing, removing stop words, and lemmatizing the text data.

- **Removing Special Characters and Punctuation Marks**
  Special characters and punctuation marks are removed from each question and answer in the dataset, using the regular expression (re) library in Python. The purpose of this step is to eliminate noise from the text data, which could interfere with the analysis.
- **Lowercasing**
  Text data in each question and answer of the dataset is converted to lowercase, to ensure consistency and reduce the number of unique words in the dataset that could result from different casing. This is an important step in preparing the data for analysis, as it standardizes the text.
- **Tokenization**
  Questions and answers are then tokenized using the nltk library in Python. This step can help with counting and analyzing the occurrence of individual words in the dataset.
- **Stopword Removal**
  Stopwords are common words such as "the," "and," and "is", that carry little meaning. In this step, stopwords are removed from the tokenized data using the nltk library. This technique can help reduce the size of the dataset and improve the accuracy of the analysis by eliminating noise and irrelevant information.
- **Lemmatization**
  In the final step of pre-processing, the tokenized data is lemmatized, after performing POS tagging, using WordNetLemmatizer from the nltk library. Lemmatization reduces tokens to their base form or lemma, which can help in reducing the number of unique words in the dataset and improving the accuracy of the analysis.

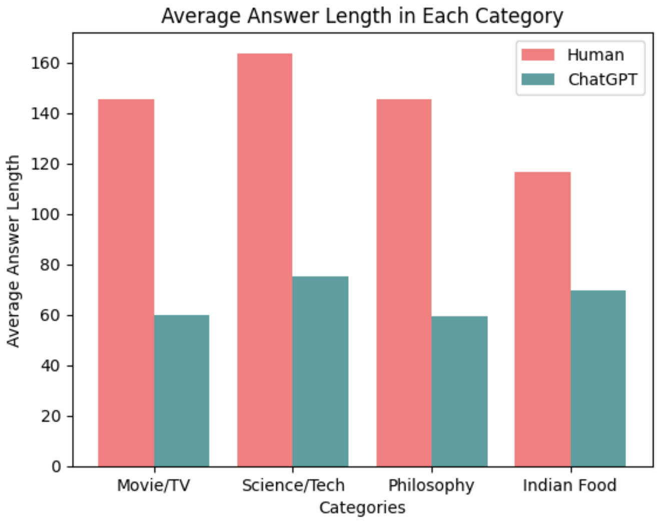


	Question	Human Answer	ChatGPT Answer	Category
0	[best, example, show, tell, movie]	[remember, incredibles, 2004, buddy, flashback...	[one, example, show, tell, movie, famous, open...	Movie/TV
1	[opinion, realistic, scene, moment, line, comi...	[line, avenger, 2012, kill, know, try, get, lo...	[ai, language, model, personal, opinion, belie...	Movie/TV
2	[awesome, movie, fact, detail]	[watched, incredibles, know, tell, early, movi...	[1, matrix, code, see, scroll, screen, actuall...	Movie/TV
3	[ever, make, entire, movie, theater, laugh]	[wife, watch, x, men, day, future, past, scene...	[ai, language, model, physically, interacted, ...	Movie/TV
4	[think, daniel, radcliffe, make, mistake, work...	[short, get, arrest, imprisoned, life, give, d...	[ai, language, model, opinion, however, worth,...	Movie/TV
...	...	...	...	...
995	[indian, dish, kerala, call, kadala, curry, k,...	[kadala, malayali, word, chickpea, also, know,...	[ai, language, model, access, every, kadala, c...	Indian Food
996	[eat, curry, dish, acid, reflux, gerd]	[name, carrie, herrera, would, like, share, te...	[depends, individual, type, curry, dish, peopl...	Indian Food
997	[many, restaurant, serve, curry, japanese, sty...	[yup, time, japanese, style, curry, top, choic...	[curry, japanese, style, also, know, kare, rai...	Indian Food
998	[feed, dog, red, curry]	[short, answer, dog, digestive, tract, make, s...	[ai, language, model, personal, preference, re...	Indian Food
999	[dog, eat, curry]	[dog, eat, curry, curry, spicy, dish, also, co...	[ai, language, model, personal, preference, fe...	Indian Food

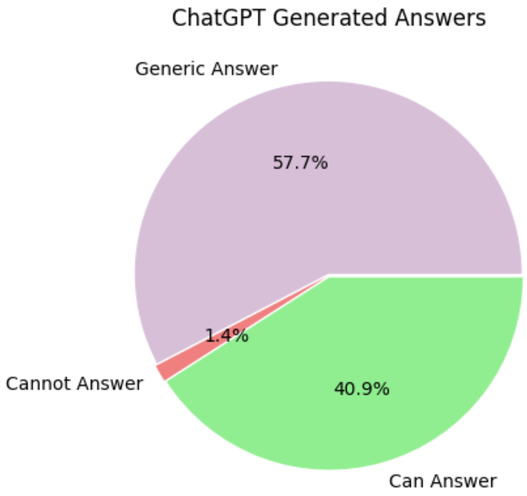
1000 rows × 4 columns

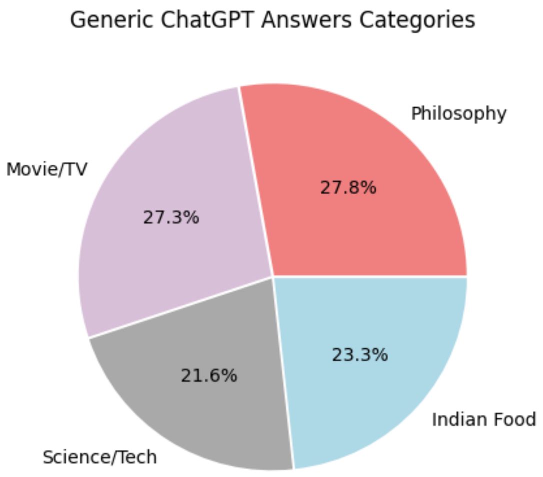
EXPLORATORY DATA ANALYSIS

- Retaining stop words may be advantageous in certain circumstances, such as authorship attribution or language identification, when their use might offer helpful hints about the language or writing style of the text. Contrarily, applications like sentiment analysis, text categorization, and topic modeling, where the emphasis is on the important words or phrases in a text, can benefit from the removal of stopwords. Before eliminating stopwords in situations where their significance is unknown, it is preferable to compare their usage. Hence, the top 20 stopwords that frequently appeared in ChatGPT and human answers are plotted. Conclusion can be reached by a statistical comparison of evaluation measures whether or not to include stopwords in the model.
- Average length of the human and ChatGPT generated answers is calculated across the four domains: Movie/TV, Science/Tech, Philosophy, and Indian Food. This analysis provides insights into the complexity of the answers generated by the model and humans for each category. To visualize the results, a bar graph is plotted, which displays the average length of answers for each category.

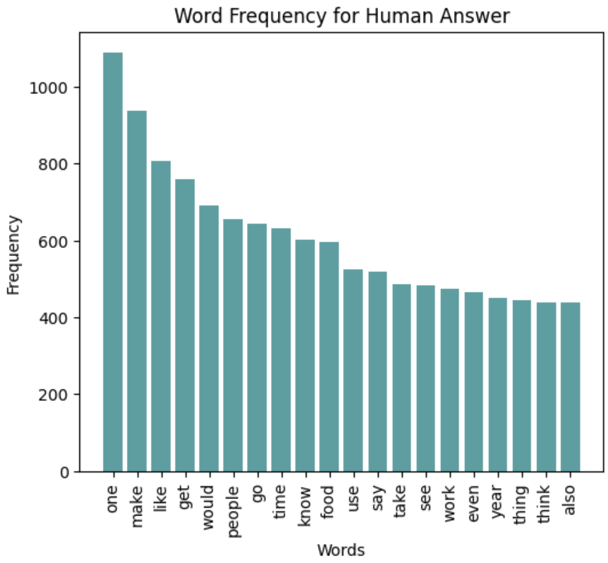


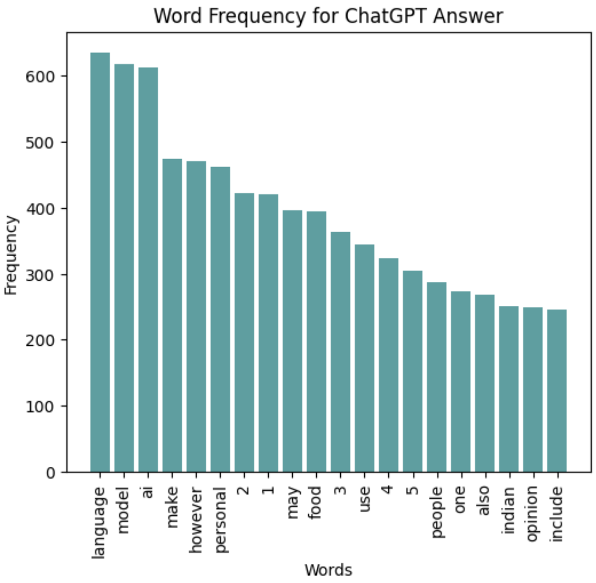
- Ratio of questions that ChatGPT cannot answer are determined. This is done by finding the words like "ai", "language" and "model". This is done because for the questions that ChatGPT is unable to provide an answer to, it starts with "As an AI language model…"



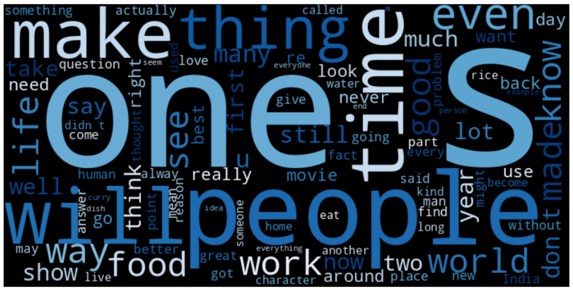


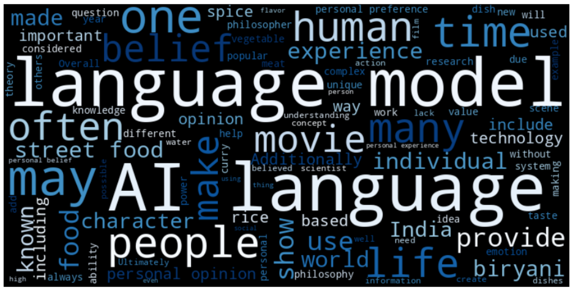
- Frequency dictionaries are constructed for human and ChatGPT answers which contain the count of each word in the corresponding set of answers, providing valuable insight into the most commonly used words and their frequency in the data. A bar graph is plotted for an easy visual comparison of the most commonly used words in the human and ChatGPT answers. The first chart shows the 20 most frequent words used in the human answers, while the second chart shows the 20 most frequent words used in ChatGPT answers, where the height represents the occurrence.



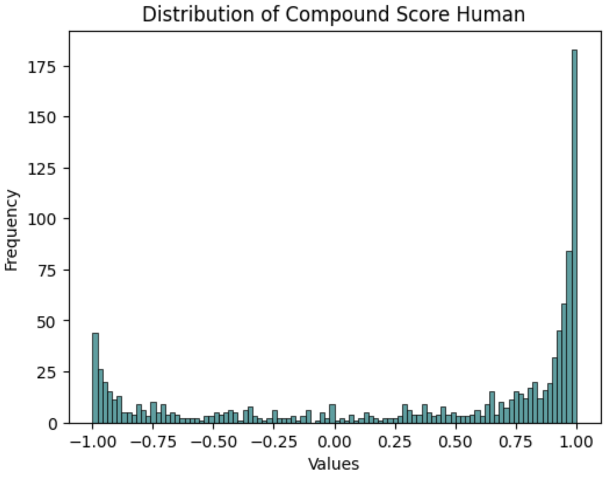


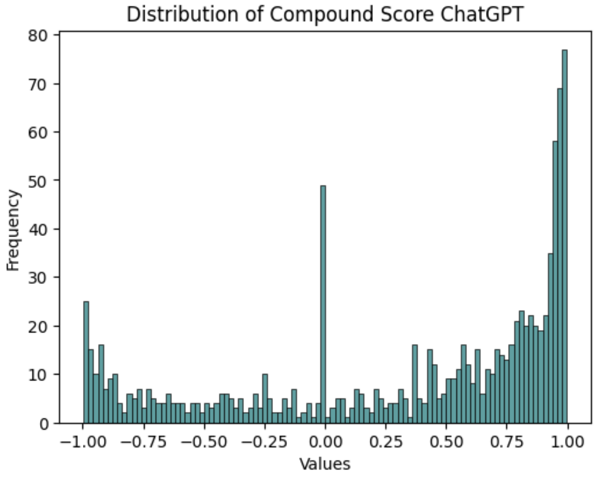
- Word Cloud is generated to visualize most frequent words in human and ChatGPT generated answers which can be useful for identifying patterns and trends in the data.





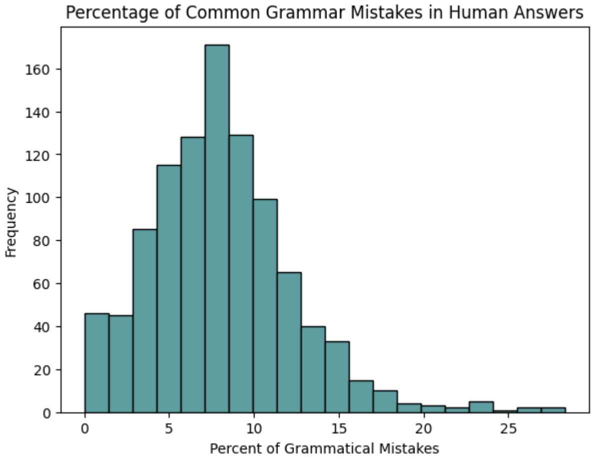
- Sentiment analysis is performed on the tokenized data, using SentimentIntensityAnalyzer function. The polarity_scores method is applied to each answer in the dataset, generating sentiment scores for positivity, negativity, neutrality, and weighted compound. The scores for each category are stored in separate lists to enable further analysis. Histograms are then plotted for each of the scores, separately for human and ChatGPT answers, providing a visualization of the distribution of sentiment scores across the dataset.

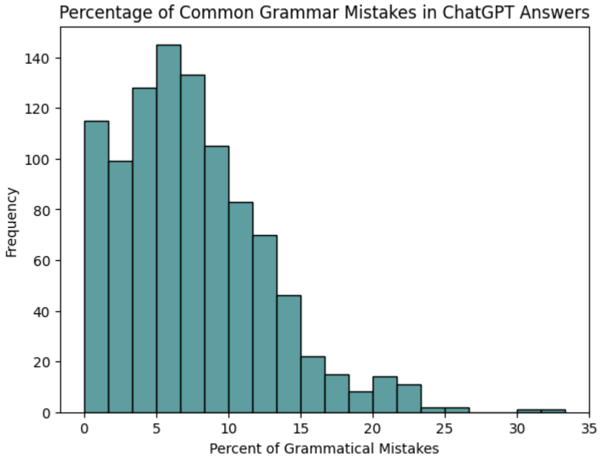




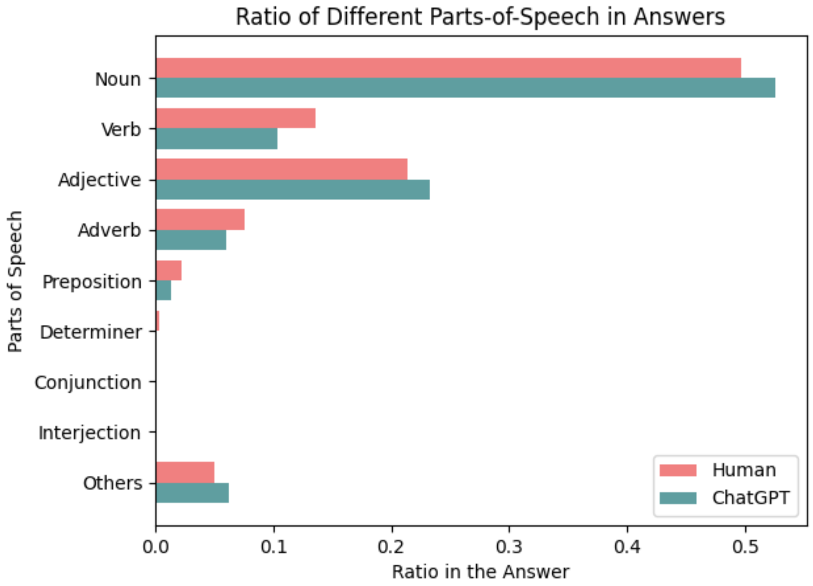
FEATURE ENGINEERING

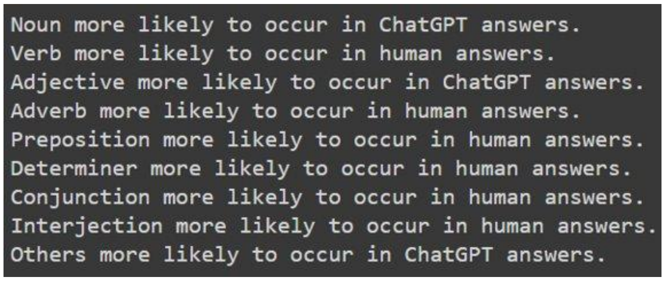
- The data is modified by concatenating the human and ChatGPT answers into one column, alongside their corresponding questions, and a target column is appended which can be used for classification. The target label 0 represents a ChatGPT answer and 1 represents a human answer. Features are extracted from the answer tokens by constructing a Term Frequency-Inverse Document Frequency matrix, which represents the importance of each term in each document, taking into account how often it appears in that document as well as how rare it is across all documents.
- In English, there are some pairs of parts-of-speech that do not occur together, such as: Preposition - Adverb, Verb - Adverb, Adverb - Adjective, Adjective - Adjective, Preposition - Preposition, Pronoun - Pronoun, Adverb - Adverb, Adjective - Pronoun, Interjection - Conjunction, etc. Ratio of such grammatical mistakes is found for both human and ChatGPT answers. It is noticed that mistakes are more likely to occur in the human answers.





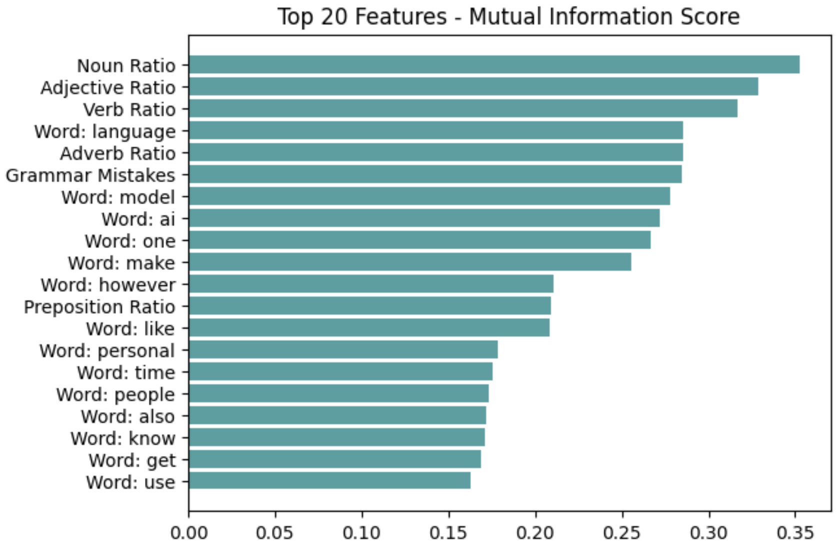
- To gain insights into the differences between human and ChatGPT-generated answers, frequency of different parts-of-speech (POS) in the answers can be analyzed. It is observed that certain parts-of-speech may be more likely to appear in human answers, while others are more common in ChatGPT generated answers. To compare the frequency of each POS in human and ChatGPT generated answers, the ratio of each POS is calculated for both sets of answers. By analyzing the ratios for each type of POS, it is identified whether it is more likely to occur in human answers or ChatGPT-generated answers.





```
Noun more likely to occur in ChatGPT answers.
Verb more likely to occur in human answers.
Adjective more likely to occur in ChatGPT answers.
Adverb more likely to occur in human answers.
Preposition more likely to occur in human answers.
Determiner more likely to occur in human answers.
Conjunction more likely to occur in human answers.
Interjection more likely to occur in human answers.
Others more likely to occur in ChatGPT answers.
```

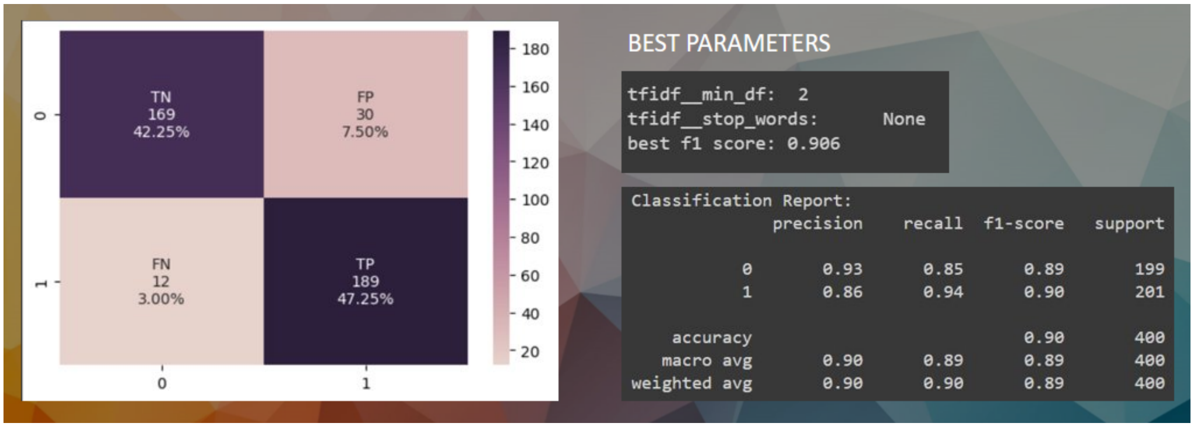
- Mutual information test is performed to identify the top features that contribute to the classification. The resulting bar graph displays the top 20 features along with their mutual information score.



MODEL IMPLEMENTATION AND EVALUATION

The additional features like length, percent of grammar mistakes, and ratio of various parts of speech, along with the TF-IDF matrix are given as the input to train our classification models.

❖ Support Vector Machine

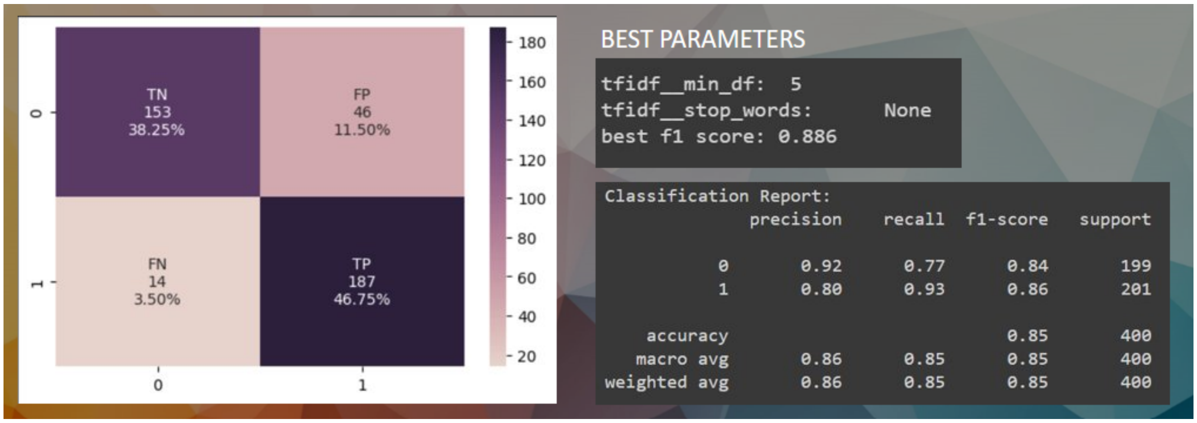


BEST PARAMETERS

```
tfidf__min_df:  2
tfidf__stop_words:      None
best f1 score: 0.906
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.85      0.89       199
           1       0.86      0.94      0.90       201

    accuracy                           0.90       400
   macro avg       0.90      0.89      0.89       400
weighted avg       0.90      0.90      0.89       400
```

❖ Naive Bayes

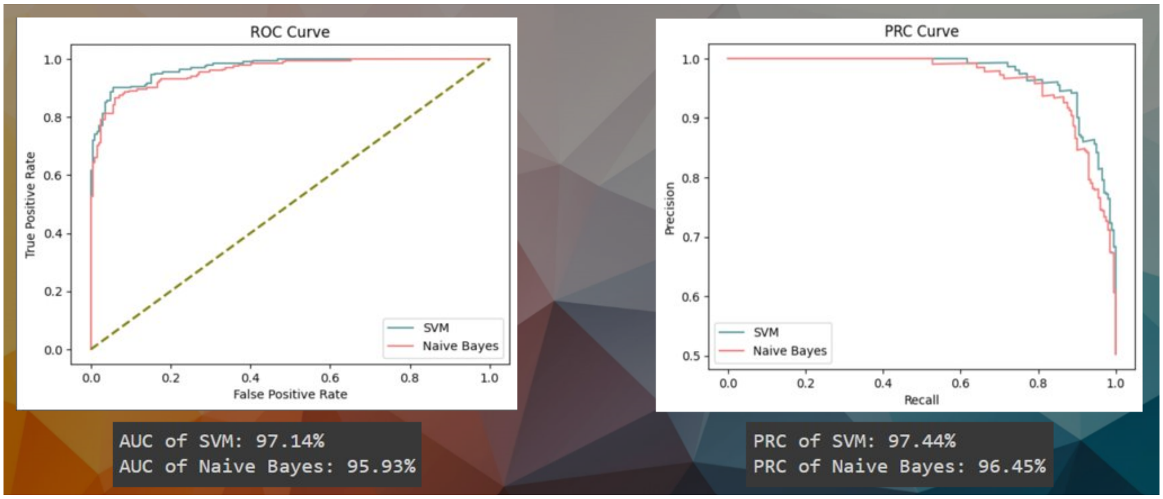


BEST PARAMETERS

```
tfidf__min_df:  5
tfidf__stop_words:      None
best f1 score: 0.886
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.77      0.84       199
           1       0.80      0.93      0.86       201

    accuracy                           0.85       400
   macro avg       0.86      0.85      0.85       400
weighted avg       0.86      0.85      0.85       400
```

The accuracy of SVM was found to be 90%, while Naive Bayes had an accuracy of 85%. SVM outperformed Naive Bayes in terms of AUC, with a score of 97.14%, compared to Naive Bayes' score of 95.93%. Additionally, SVM had a higher PRC score of 97.44% compared to Naive Bayes' score of 96.45%.



```
AUC of SVM: 97.14%
AUC of Naive Bayes: 95.93%
```

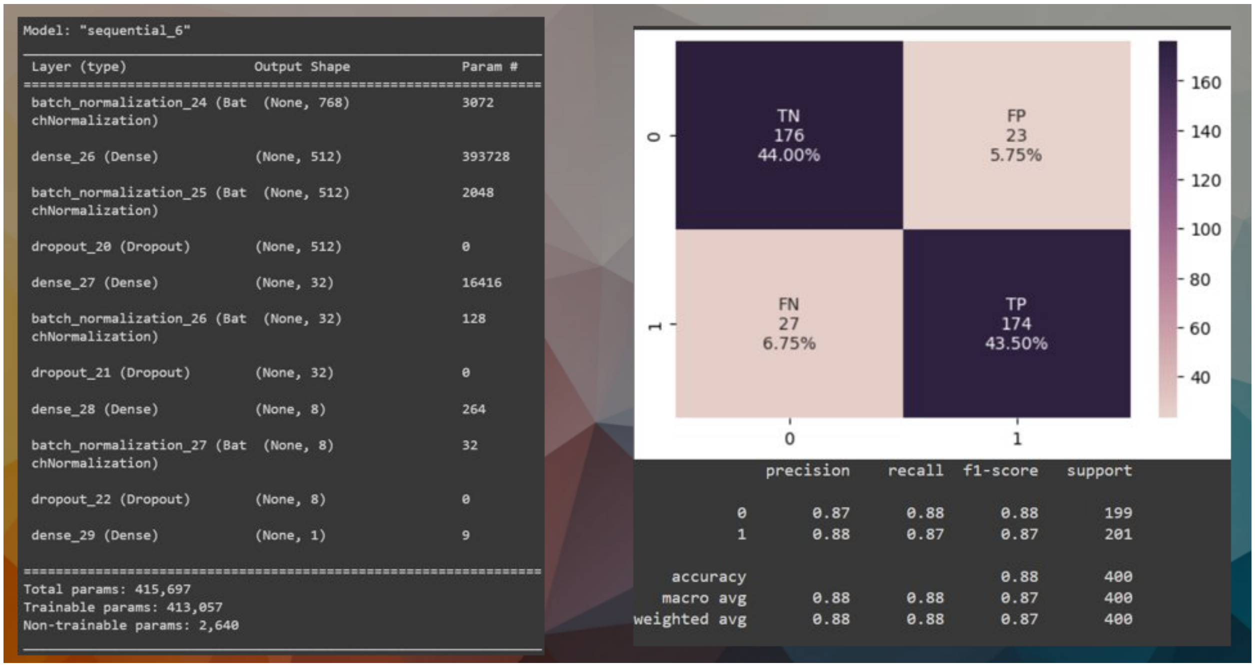
```
PRC of SVM: 97.44%
PRC of Naive Bayes: 96.45%
```

- ❖ Neural Networks:

The TFIDF is not capable of capturing semantic meaning from the document. So, we implemented the sentence encoding. **SentenceTransformers** is a Python framework for state-of-the-art sentence, text and image embeddings. The initial work is described in paper [Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks]

The framework is used to compute text embedding in more than 100 languages. This framework is intended to be used as a sentence and short paragraph encoder. Given an input text, it outputs a vector which captures the semantic information. The sentence vector may be used for information retrieval, clustering or sentence similarity tasks

In this project the embeddings are utilized for fine tuning our model for classification and determining the similarity between human and ChatGPT answers. Model used for embedding is 'all-MiniLM-L6-v2', it maps sentences from English to a 384-dimensional dense vector space.



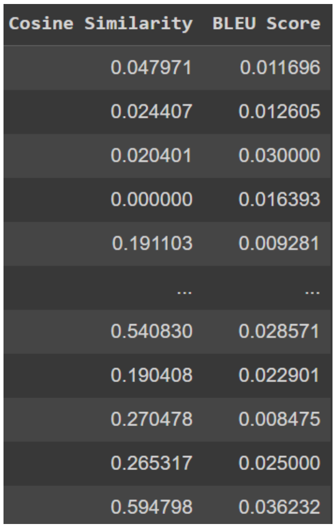
```
Model: "sequential_6"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 batch_normalization_24 (Bat  (None, 768)              3072
 chNormalization)

 dense_26 (Dense)            (None, 512)               393728

 batch_normalization_25 (Bat  (None, 512)              2048
 chNormalization)

 dropout_20 (Dropout)        (None, 512)               0

 dense_27 (Dense)            (None, 32)                16416

 batch_normalization_26 (Bat  (None, 32)               128
 chNormalization)

 dropout_21 (Dropout)        (None, 32)                0

 dense_28 (Dense)            (None, 8)                 264

 batch_normalization_27 (Bat  (None, 8)                32
 chNormalization)

 dropout_22 (Dropout)        (None, 8)                 0

 dense_29 (Dense)            (None, 1)                 9

=================================================================
Total params: 415,697
Trainable params: 413,057
Non-trainable params: 2,640
_________________________________________________________________
```

```
              precision    recall  f1-score   support

           0       0.87      0.88      0.88       199
           1       0.88      0.87      0.87       201

    accuracy                           0.88       400
   macro avg       0.88      0.88      0.87       400
weighted avg       0.88      0.88      0.87       400
```

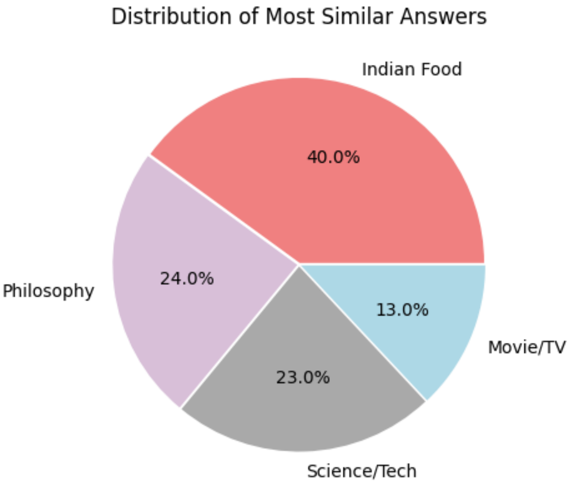
Our neural network model has 3 hidden layers. 'Sigmoid' is used as the output activation function so that the model can perform binary classification. Batch normalization, L2 regularization and dropout are the techniques implemented to avoid overfitting. After training and testing, we were able to achieve an accuracy of 88% and F1-score of 0.87.

SIMILARITY ANALYSIS

The similarity between ChatGPT-generated and human answers was analyzed using two metrics: cosine similarity and BLEU score. It is inferred that the BLEU score may not provide an accurate analysis of similarity between answers as it only captures information about the overlapping n-grams.

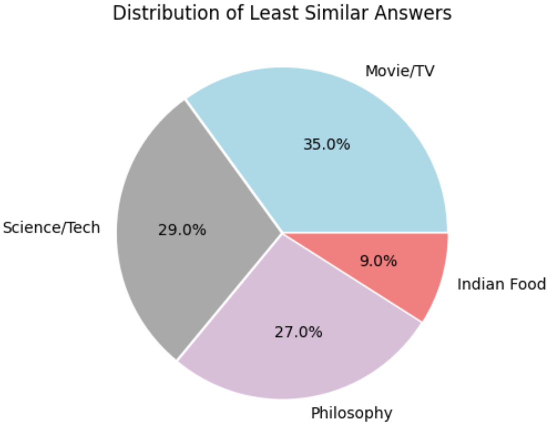


Cosine Similarity	BLEU Score
0.047971	0.011696
0.024407	0.012605
0.020401	0.030000
0.000000	0.016393
0.191103	0.009281
...	...
0.540830	0.028571
0.190408	0.022901
0.270478	0.008475
0.265317	0.025000
0.594798	0.036232

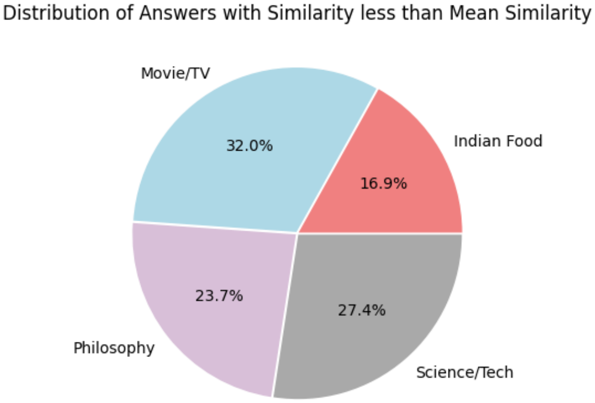


Distribution of Most Similar Answers

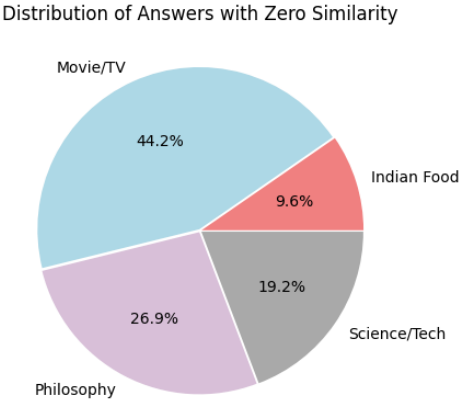
Due to this reason, cosine similarity is preferred. It is used to find the domains where chatgpt had most similar answers as humans. Our analysis revealed that ChatGPT was most successful in mimicking human answers in the Indian Food category, while it performed the least in the Movie/TV category.



Distribution of Least Similar Answers



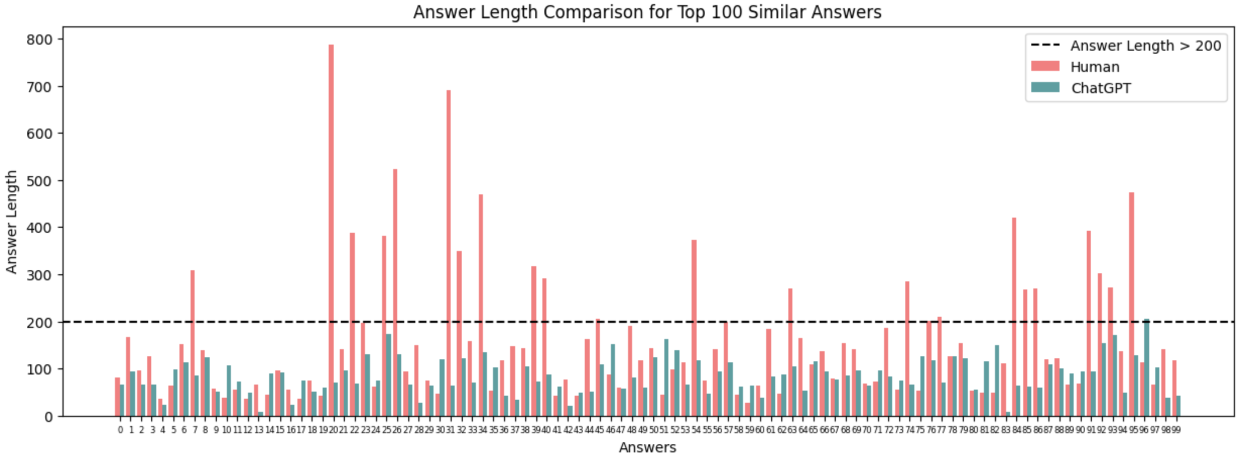
Distribution of Answers with Similarity less than Mean Similarity

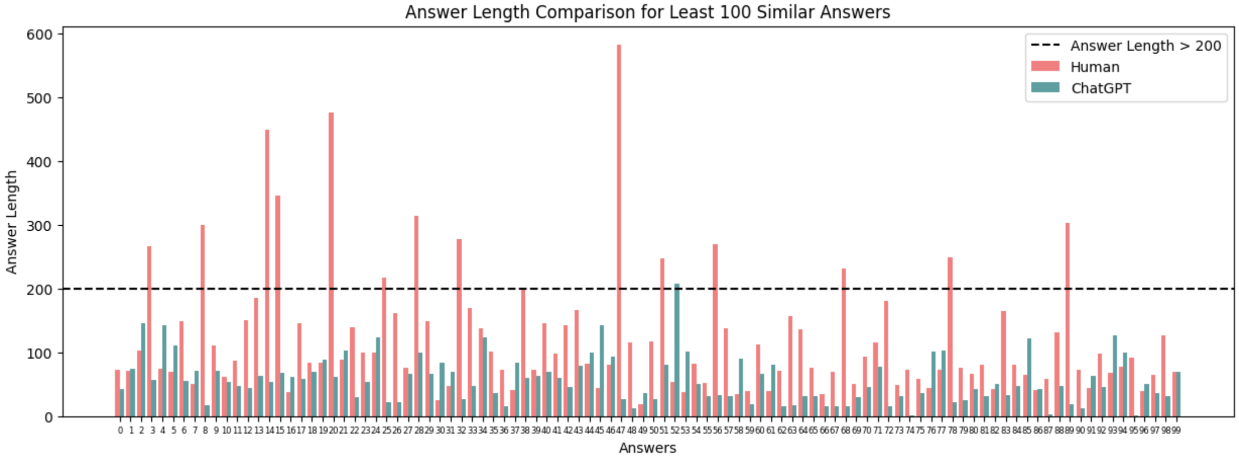


Distribution of Answers with Zero Similarity

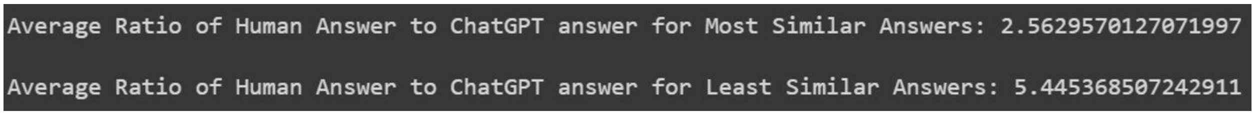
CAPABILITIES AND LIMITATIONS OF CHATGPT

1. Further analysis is done on the most similar and the least similar answers. The below graphs display the length of each answer for the most similar and least similar answers. It is difficult to draw conclusions only by looking at these graphs.



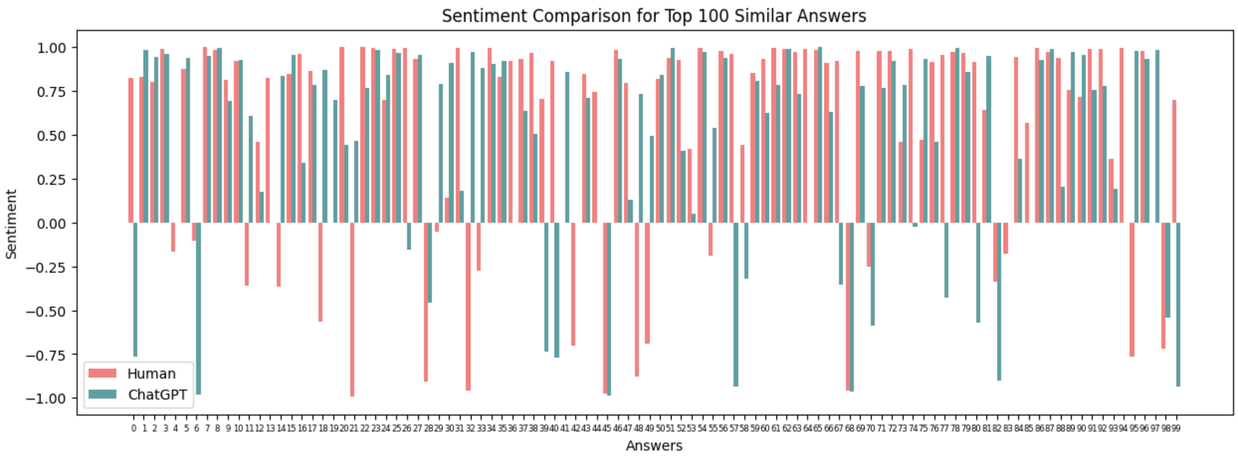


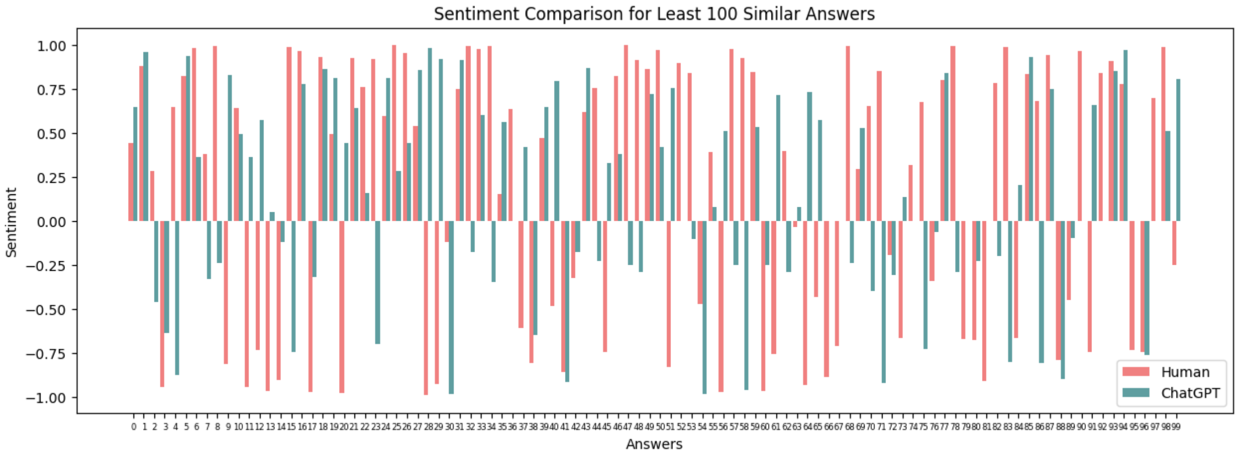
Upon analyzing the average human to ChatGPT ratio, we observed that for the most similar answers, the ratio of answers was approximately 2.5. In contrast, for the least similar answers, the ratio was almost double, approximately 5.5.



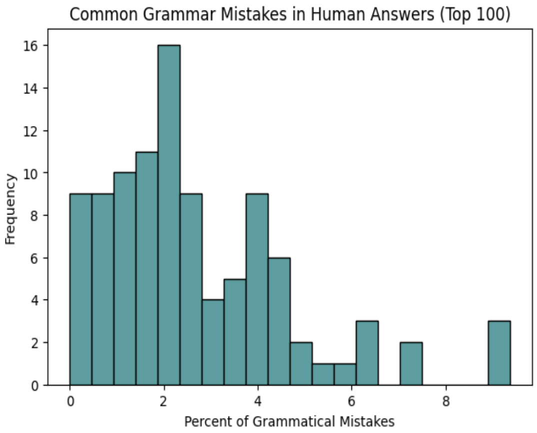
```
Average Ratio of Human Answer to ChatGPT answer for Most Similar Answers: 2.5629570127071997

Average Ratio of Human Answer to ChatGPT answer for Least Similar Answers: 5.445368507242911
```

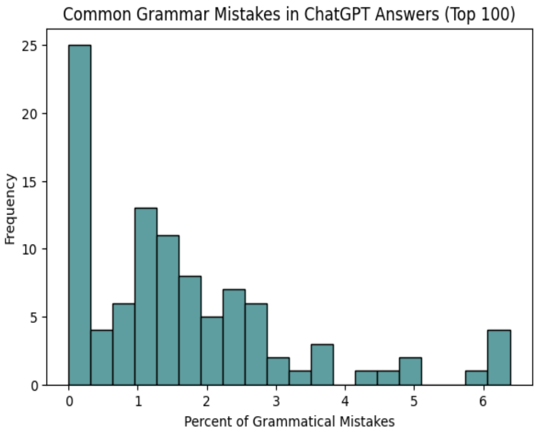
2. Sentiments of the most and least similar answers were also examined. It was observed that the majority of the most similar answers had a positive sentiment. In contrast, there was no clear pattern in the sentiments of the least similar answers.

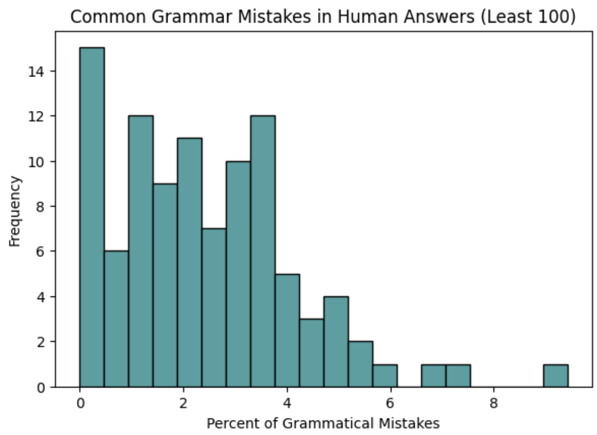


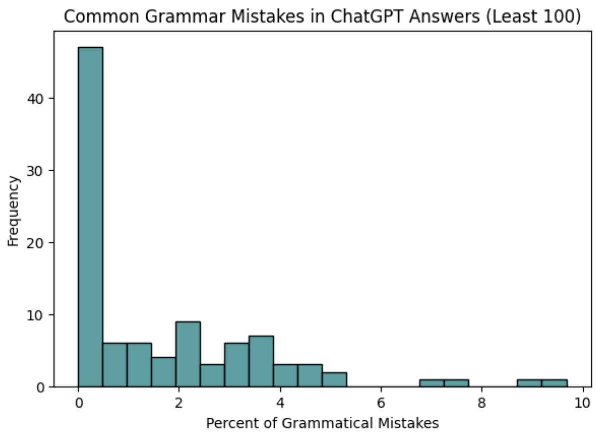


3. The percentage of common grammatical mistakes was checked for human and ChatGPT answers. The following graphs show that overall, humans tend to make more grammatical mistakes. However, in the least similar answers, ChatGPT had a very low ratio of grammatical mistakes.









CONCLUSION

In summary, we have completed three tasks in our project.

- ❖ Developed a classifier to distinguish between human and ChatGPT answers, using SVM, Naïve Bayes, and neural networks. SVM performed the best among them.
- ❖ Assessed the similarity between human and ChatGPT answers using cosine similarity and found that ChatGPT provided more similar answers to humans for fact-based questions and least similar answers for opinion-based questions.
- ❖ Analyzed the limitations and capabilities of ChatGPT, and got an interesting insight that for questions with multiple correct answers, humans and ChatGPT may not provide a similar answer, but they can both be correct.

Overall, our project highlights the strengths and limitations of ChatGPT in generating answers that are similar to human answers. While ChatGPT has the potential to be a useful tool in many domains, it is important to consider its limitations and continue to improve the model to achieve even better performance.

FUTURE WORK

Future work includes expanding the scope of the study to include a larger number of domains, enabling a more comprehensive understanding of the similarity between human and ChatGPT-generated answers and identifying opportunities to enhance the performance of the ChatGPT model. Diversity of ChatGPT answers and human answers can be assessed. Evaluating which of the two types of answers are more diverse can offer insights into the quality and effectiveness of the answer generation process. A higher level of diversity in human answers could indicate a greater understanding of the complexities of the domain, leading to a wider range of perspectives and solutions. Conversely, if ChatGPT-generated answers demonstrate greater diversity, it could suggest that the model can produce a wider range of responses and has a better grasp of the domain's underlying concepts.