GenFiction: Analysing Fiction Written by LLMs

This project aims to understand the linguistic differences between fiction written by humans and fiction written by LLMs. It focuses solely on flash fiction, ie fictional stories between 500 to 1000 words.

Dataset:

- A novel dataset has been created for this project
- For human-written fiction, scrapy library has been used to scrape over 1000 stories from flashfictiononline.com and everydayfiction.com. This dataset is present in real_fiction_data.csv in the datasets folder. It has the following attributes:
  - url: url from which data has been scraped
  - title: the title of the story
  - author: the name of the author
  - content: the content of the fiction story
- For AI-written fiction, the dataset has been manually created by passing in writing prompts to ChatGPT. This dataset is present in generated_fiction_data.csv and contains 200 stories.
- To create a balanced dataset, 200 human-written stories and 200 AI-written stories have been used to create a final dataset containing 400 stories. This final dataset, containing the articles and target label isReal, is present in shuffled_data_new.csv
- The dataset along with all the extracted features is present in data_with_all_features.csv

Feature Extraction:

- The features being analysed in this project broadly come under 3 categories – Complexity features, Stylistic features, and Psychology features
- Stanza library has been used for the extraction of both complexity and stylistic features.
- Empath library has been used for the extraction of psychology features

Features:

- A total of 25 features have been extracted and used for this project
- The complexity features being used in this project are:
  - Average Sentence Length: Average no. of words per sentence
  - Average Complex Words: Ratio of no. of complex words (words with 3 syllables or more) to total no. of words
  - Average syllables per word: Average no. of syllables in each word
  - Gunning Fog Readability Index
  - Flesch Reading Ease Readability Formula

- The stylistic features being used in this project are:
  - Ratio of number of nouns to total no. of words
  - Ratio of number of adjectives to total no. of words
  - Ratio of number of verbs to total no. of words
  - Ratio of no. of adverbs to total no. of words
  - Ratio of no. of conjunctions to total no. of words
  - Ratio of no. of pronouns to total no. of words
  - Ratio of no. of first/second/third person words to total no. of words
  - Ratio of no. of present/past/future tense words to total no. of words
  - Ratio of no. of active/passive words to total no. of words

- The psychology features being using in this project are:
  - No. of Hate-related words
  - No of Family-related words
  - No of Love-related words
  - No of Crime-related words
  - No of Optimistic words
  - No of Violence-related words
  - No of Sadness- related words
  - No of Emotional words
  - No of Joy- related words
  - No of words expressing negative emotion
  - No of words expressing positive emotion

Modelling:

- 2 machine learning models – Logistic regression and linear Support Vector Machine (SVM) have been used
- Accuracies have been computed for 4 different feature sets- Only complexity features, only stylistic features, only psychology features and lastly, all the features put together
- Test-train split is 30-70.
- SHAP has been used for the global interpretation of both models on all the four feature sets

Results and Discussion:

- For complexity features, the following accuracies were obtained:
  Logistic Regression Accuracy: 0.775
  Linear SVM Accuracy: 0.775
- Gunning Fog readability index exhibited the largest mean SHAP values, indicating that this feature has the largest contribution to the model predictions.
- The violin plot also showed that larger Gunning Fog readability index values showed a lower SHAP value where as smaller Gunning Fog readability index values showed higher SHAP values.
- This seems to indicate that human-written fiction has a tendency to have greater readabilty as opposed to AI-written fiction.

- For stylistic features, the following accuracies were obtained:
  Logistic Regression Accuracy: 0.7833333333333333
  Linear SVM Accuracy: 0.7666666666666667
- The past tense, present tense, first person and third person features have the greatest impact on the model predictions.
- Fiction written by GenAI is very likely to be written in third person and in past tense as opposed to human-written fiction

- For psychology features, the following accuracies were obtained:
  Logistic Regression Accuracy: 0.7333333333333333
  Linear SVM Accuracy: 0.5666666666666667
- Words indicating positive emotion, love and optimism had the greatest impact on the model
- Fiction written by GenAI had a tendency to be more overwhelmingly positive as opposed to human-written fiction

- Lastly, when all the features were considered, the following accuracies were obtained:
  Logistic Regression Accuracy: 0.8333333333333334
  Linear SVM Accuracy: 0.875

- Interestingly, when all features were taken into account, complexity features like Gunning Fog Readability Index, Flesch Readability Index and average number of words per sentence had the greatest impact.
- There is strong evidence to suggest that for human-written fiction, the Gunning Fog readability index is lower, ie the text is more readable. Contrary to this, however, human written fiction had greater sentence lengths. Gunning Fog index is derived from both sentence length as well number of complex words. Hence, this points to the fact that human-written fiction tends to use lesser complex words as opposed to GenAI fiction.

To conclude, the high accuracy (87.5%) obtained indicates that there is still a vast difference between human-written fiction and AI-generated fiction when it comes to lingustic features. These differences are mainly in the structure and complexity of the text, as well as the tense and point-of-view words used.

Limitations:

- The major limitation of this project is the size of the dataset used. 400 texts have been used for this task, which may be insufficient and can lead to overfitting. It may also be beneficial to use k-fold cross validation as opposed to a 70-30 train-test split.