Human vs. AI Text Classification with Deep Learning - Final Project

Orel Dayan 209452093
Sagi Azulay 207544230
Roy Asraf 302211958

Abstract

This project focuses on classifying texts as written by humans or generated by AI using deep learning models. Features like word usage, punctuation, synonym percentage, and text length are extracted into Logistic Regression and CNN models. The models achieve over 99% accuracy on a Kaggle dataset of 10,000 human-written and 10,000 AI-generated texts. The CNN model slightly outperforms logistic regression.

1 Introduction

AI text generation systems have become very advanced, producing remarkably human-like writing. The goal of this project was to build a classification model capable of distinguishing between human-generated and AI-generated texts based on various features extracted from the text data.

We conducted preprocessing on the text data, which involved several key steps to prepare it for classification analysis. These steps included tokenization, removal of stopwords and linking words, sentiment analysis, and extraction of various features such as punctuation count, linking words count, and text length. Additionally, binary features were introduced to represent the presence or absence of the top 500 most common words in the dataset (word_features) With logistic regression, progressed to CNN of the conv1d type.

This report will present the results, key learnings, and methodologies employed in the research process.

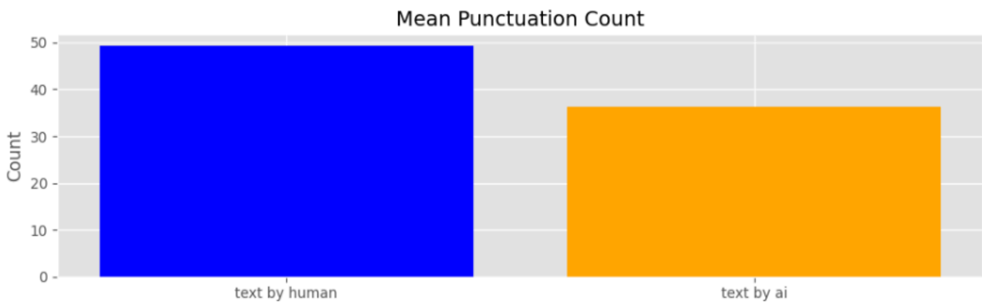
In the example below we check different values of common words and determine the effect on the results.

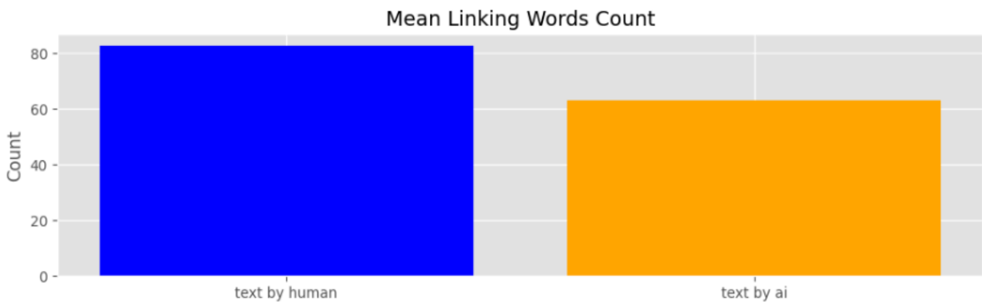
Common words	test loss	train loss	accuracy
50	8.6972%	8.7519%	96.633333%
100	3.11922%	3.196%	99%
200	1.458%	0.93%	99.5%
500	0.74248%	0.3398%	99.766%

Out[13]:

	text	generated	punctuation_count	linking_words_count	length_text	word_features	punctuation_count_percentage	linking_words_percentage
0	Cars. Cars have been around since they became ...	0.0	75	131	657	[1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, ...	11.398176	19.908815
1	Transportation is a large necessity in most co...	0.0	64	108	526	[0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, ...	12.167300	20.532319
2	"America's love affair with it's vehicles seem...	0.0	101	162	842	[1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, ...	11.995249	19.239905
3	How often do you ride in a car? Do you drive a...	0.0	124	133	805	[1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, ...	15.403727	16.521739
4	Cars are a wonderful thing. They are perhaps o...	0.0	110	155	967	[1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, ...	11.351909	15.995872
...	...	...	...	...	...	...	...	...
26090	The use of renewable energy sources is an impo...	1.0	20	44	316	[0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, ...	6.329114	13.924051
26091	High school sports are often a source of pride...	1.0	24	46	349	[1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, ...	6.876791	13.180516
26092	The beauty of nature can be seen in the cycle ...	1.0	26	42	355	[1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, ...	7.323944	11.830986
26093	The impact of air pollution on human health is...	1.0	26	47	377	[1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, ...	6.896552	12.466844
26094	It is often said that the best things in life ...	1.0	26	41	300	[1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, ...	8.666667	13.666667

20000 rows × 8 columns





2 Related Work and Required Background

2.1 Related Work

Several studies have explored text classification, but the increasing sophistication of AI-generated texts presents new challenges. Techniques ranging from simple Logistic Regression to complex neural networks have been employed. Familiarity with natural language processing (NLP), sentiment analysis, and deep learning fundamentals, including logistic regression and neural networks, is essential for understanding this study.

2.2 Required background

Logistic regression is a statistical method used for binary classification tasks. It models the probability that a given input belongs to a particular class using a logistic function. In the context of this project, logistic regression is employed to classify text samples as human-generated or AI-generated based on extracted features.

CNNs are a type of deep learning model commonly used for image recognition, but they can also be applied to sequential data like text. Convolutional layers are the building blocks of CNNs. These layers consist of filters or kernels that slide over the input data (e.g., an image or a sequence of words) to perform convolution operations. In a CNN architecture, conv1 specifically refers to the first convolutional layer. This layer typically operates directly on the raw input data or the output of an initial embedding layer (in the case of text data). Conv1 plays a crucial role in capturing low-level features from the input, which are then passed on to subsequent layers for further processing. During the convolution operation in conv1, the filter slides across the input data, and at each position, it computes the element-wise multiplication between the filter and the corresponding patch of the input data. The results of these multiplications are summed up to produce a single output value for that position. After the convolution operation, the resulting feature maps often undergo activation functions to introduce non-linearity. Additionally, pooling layers may be applied to reduce the spatial dimensions of the feature maps while retaining the most important information. In text classification tasks, conv1 is typically used to extract local patterns or features from sequences of words (or word embeddings). By sliding filters over the input text, conv1 can learn to detect important textual patterns that are relevant to the classification task.

In text data analysis for deep learning, recurrent neural networks (RNNs) have traditionally been favored due to their ability to capture sequential patterns inherent in text. However, recent studies have shown that 1D convolutional neural networks (CNNs) can also effectively capture sequence information. Furthermore, CNNs with 1D convolution require fewer parameters and train faster compared to RNNs.

3 Project Description

text
'cars. cars have been around since they became famous in the 1900s, when henry ford created and built the first modelt. cars have played a major role in our every day lives since then. but now, people are starting to question if limiting car usage would be a good thing. to me, limiting the use of cars might be a good thing to do. in like matter of this, article, "in german suburb, life goes on without cars," by elizabeth rosenthal states, how automobiles are the linchpin of suburbs, where middle class families from either shanghai or chicago tend to make their homes. experts say how this is a huge impediment to current efforts to reduce greenhouse gas emissions from tailpipe. passenger cars are responsible for 12 percent of greenhouse gas emissions in europe...and up to 50 percent in some carintensive areas in the united states. cars are the main reason for the greenhouse gas emissions because of a lot of people driving them around all the time getting where they need to go. article, "paris bans driving due to smog," by robert duffer says, how paris, after days of nearrecord pollution, enforced a partial driving ban to clear the air of the global city. it also says, how on monday, motorist with evennumbered license plates were ordered to leave their cars at home or be fined a 22euro fine 31. the same order would be applied to oddnumbered plates the following day. cars are the reason for polluting entire cities like paris. this shows how bad cars can be because, of all the pollution that they can cause to an entire city. likewise, in the article, "carfree day is spinning into a big hit in bogota," by andrew selsky says, how programs that\'s set to spread to other countries, millions of columbians hiked, biked, skated, or took the bus to work during a carfree day, leaving streets of this capital city eerily devoid of traffic jams. it was the third straight year cars have been banned with only buses and taxis permitted for the day without cars in the capital city of 7 million. people like the idea of having carfree days because, it allows them to lesson the pollution that cars put out of their exhaust from people driving all the time. the article also tells how parks and sports centers have bustled throughout the city uneven, pitted sidewalks have been replaced by broad, smooth sidewalks rushhour restrictions have dramatically cut traffic and new restaurants and upscale shopping districts have cropped up. having no cars has been good for the country of columbia because, it has aloud them to repair things that have needed repairs for a long time, traffic jams have gone down, and restaurants and shopping districts have popped up, all due to the fact of having less cars around. in conclusion, the use of less cars and having carfree days, have had a big impact on the environment of cities because, it is cutting down the air pollution that the cars have majorly polluted, it has aloud countries like columbia to repair sidewalks, and cut down traffic jams. limiting the use of cars would be a good thing for america. so we should limit the use of cars by maybe riding a bike, or maybe walking somewhere that isn\'t that far from you and doesn\'t need the use of a car to get you there. to me, limiting the use of cars might be a good thing to do.'

length_text	word_features	synonym_percentage	punctuation_count_percentage	linking_words_percentage
300	[1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, ...	69.000000	8.666667	13.666667

```
model = Sequential()
model.add(Conv1D(filters=64, kernel_size=3, activation='relu', input_shape=(X_train_cnn.shape[1], 1)))
model.add(MaxPooling1D(pool_size=2))
model.add(Conv1D(filters=128, kernel_size=3, activation='relu'))
model.add(MaxPooling1D(pool_size=2))
model.add(Flatten())
```

The final approach consists of preprocessing the data, extracting relevant features, and training both Logistic Regression and Conval Neural Network models. The text data was tokenized, and preprocessing steps like removing stopwords, non-alphabetic characters, and stemming were applied. Also, we removed duplicated texts. The features were: the 500 most common words, linking words count in percentage, count_synonyms in percentage, and panctuation_count in percentage and text length.

Logistic Regression Approach

In the logistic regression approach, a logistic regression classifier is trained on the extracted features to predict the class labels of text samples. The model's performance is evaluated using metrics such as accuracy, logistic loss, and a classification report containing precision, recall, and F1-score for each class.

CNN Approach

The CNN approach reshapes the input data to fit the architecture of the CNN model. The CNN model consists of convolutional layers followed by max-pooling layers and fully connected layers with dropout regularization. The model is trained on the preprocessed text data and evaluated on both training and validation sets. Accuracy and error rates are calculated to assess the model's performance.

4 Experiments/Simulation Results

The data was split 70/30 into train and test sets. 10-fold cross-validation was used during training.

Logistic Regression Results

Accuracy: 99.77%
Train
Loss: 0.0034Test Loss: 0.0074
Classification Report:
Precision, recall, and F1-score for both human-generated and AI-generated classes.
Confusion Matrix: Visual representation of model performance.

CNN Results

Train Error: 0.02%
Validation Error: 0.15%
Training History Plot: Visualization of accuracy on training and validation sets across 10 epochs.

```
Epoch 1/10
438/438 [==============================] - 18s 38ms/step - loss: 0.0714 - accuracy: 0.9721 - val_loss: 0.0182 - val_accuracy: 0.
9962
Epoch 2/10
438/438 [==============================] - 17s 38ms/step - loss: 0.0221 - accuracy: 0.9924 - val_loss: 0.0117 - val_accuracy: 0.
9953
Epoch 3/10
438/438 [==============================] - 16s 38ms/step - loss: 0.0133 - accuracy: 0.9951 - val_loss: 0.0078 - val_accuracy: 0.
9975
Epoch 4/10
438/438 [==============================] - 17s 39ms/step - loss: 0.0077 - accuracy: 0.9975 - val_loss: 0.0076 - val_accuracy: 0.
9980
Epoch 5/10
438/438 [==============================] - 17s 38ms/step - loss: 0.0073 - accuracy: 0.9973 - val_loss: 0.0061 - val_accuracy: 0.
9983
Epoch 6/10
438/438 [==============================] - 17s 39ms/step - loss: 0.0060 - accuracy: 0.9979 - val_loss: 0.0083 - val_accuracy: 0.
9972
Epoch 7/10
438/438 [==============================] - 17s 40ms/step - loss: 0.0030 - accuracy: 0.9989 - val_loss: 0.0089 - val_accuracy: 0.
9975
Epoch 8/10
438/438 [==============================] - 17s 39ms/step - loss: 0.0058 - accuracy: 0.9975 - val_loss: 0.0098 - val_accuracy: 0.
9973
Epoch 9/10
438/438 [==============================] - 17s 38ms/step - loss: 0.0047 - accuracy: 0.9983 - val_loss: 0.0072 - val_accuracy: 0.
9983
Epoch 10/10
438/438 [==============================] - 17s 38ms/step - loss: 0.0034 - accuracy: 0.9991 - val_loss: 0.0073 - val_accuracy: 0.
9985
438/438 [==============================] - 4s 9ms/step
188/188 [==============================] - 2s 9ms/step
Train Error: 0.0002142857142857224
Validation Error: 0.0014999999999999458
```

6 Previous Attempts

Recap of Assignment 2

In reviewing Assignment 2, our objective was to evaluate and compare the performance of three distinct methodologies. Specifically, we examined linear regression, softmax regression, and logistic regression techniques. Our analysis gives up the following outcomes:

- Linear regression achieved an accuracy rate of 91.2%.

- Softmax regression exhibited an accuracy rate of 98.75%.

- The most promising result was attained through logistic regression, yielding an accuracy rate of 99.76%.

These findings underscore the superior performance of logistic regression in our experimental context. Such results provide valuable insights into the efficacy of different regression techniques within the scope of our study.

Additionally, Features that had almost no effect were the most common word in each text, sentiment_score we can see that because If we compare them we will have almost the same number on average for each type of text.

7 Conclusions

Both logistic regression and CNN models demonstrate high accuracy in classifying text samples as human-generated or AI-generated. Logistic regression achieves an accuracy of 99.77%, while the CNN model achieves a slightly lower validation error of 0.15%. These results indicate the effectiveness of both approaches in distinguishing between human and AI-generated text.