Text classification concepts and model development

UC | Text Analysis

Author: Margarita Grishechkina

Problem

The key point of ‘supervised learning’ by training a model to perform text classification, for me, is understanding how the model (or algorithm) comes to a decision or why it reaches a particular conclusion rather than simply observing the results. Once we understand this process, as (Martin) discuss, we can trust the model then further discuss, interpret, predict and build a more transparent model. When creating, modifying or adding a new feature, we should understand the role it plays within the model. This allows us to experiment with variables and investigate associations, causation, hypotheses, and the relationships between inputs and outputs or variety of output types without having to read the entire text. Depending on complexity of text we can use different classifiers focusing on accuracy of predictions.

The interpretable approach allows us to understand how a model reaches a decision. As (Rudin) argues, interpretable models are designed to be transparent from the start allowing us to interpret them without any additional efforts just by simply observing their structure. Logistic regression is relevant for this approach and for the assessment task because it classifies data into one (or more) labelled categories or classes and performs well when the relationship between variables and classes is linear. Decision trees, on other hand, are relevant for non-liner interactions and allow us to to interpret the model easily by sorting data step-by-step using "yes-no-question" technique providing immediate insights into the results.

The explainable approach can’t be so easily interpreted. Models in this category are often require additional methods and technics (post-hoc explanations) to make their decisions understandable to humans. They can perform more complex text classification tasks, for example, to analyse multiply dependencies, or linguistic nuances or deep semantic understanding. Neural networks are relevant for this approach and relevant for complex text classification tasks.

As a Data Scientist working for an academic institution, my task is to build a prototype of interpretable text classification system to distinguish between essays authored by students and essays authored by large language models (LLMs). While developing prototype I aim to

demonstrate the knowledge and skills I have gained throughout the course. I plan to use logistic regression model, test potential features, compare results, patterns and other profitable findings, then choose the most relevant combination of features and describe the results in a way to understand for all kind of users.

First of all, I examined how the data was split for training and testing. In this case, the split of 69.82% for training and 30.18% for testing, as shown in the illustration below, gives us a 70/30 percentage result. This slightly shifted ratio helps improve validation of the model's performance on unseen data.

split	count	percentage
test	2100	30.18
train	4858	69.82

Figure 1, Train and test data splitting

Next, I checked the labels, with '0' for human-written text and '1' for LLM-generated text, to prepare for further analysis. Specifying these labels allows us to use a binary classifier, such as a logistic regression model, to make predictions. After that, normalization steps, like removing extra spaces and symbols, can be performed to improve the model's generalization.

Text classification model

'Token frequency' and 'bigrams' were initially considered followed by adding the 'sentencizer', that helps split a text into sentences based on solely punctuation making it more efficient than using a full parsing approach and 'lemmatizer', that converts words to their base form to simplify text for analysis. As well as that tokens were transformed tinto numerical vectors to capture their meaning for the model and .

At first, I included almost all features, using universal POS tags (Universal Dependencies contributors) and text statistics to inspect the model coefficients and to observe features relevance. The results of the model coefficients are shown in the figure below.

Model coefficients
with log odds (logit) converted to odds ratio
for improved interpretability

Target label: gpt

	feature	log odds (logit)	odds ratio
0	CCONJ_relfreq	1.034234	2.812951
1	polysyll_relfreq	0.578379	1.783145
2	ADP_relfreq	0.518355	1.679263
3	PRON_relfreq	0.371759	1.450284
4	VERB_relfreq	0.328994	1.389570
5	positive_relfreq	0.265882	1.304581
6	negative_relfreq	0.094369	1.098965
7	ADV_relfreq	0.007955	1.007986
8	DET_relfreq	-0.057530	0.944094
9	ADJ_relfreq	-0.256051	0.774102
10	PART_relfreq	-0.334158	0.715940
11	PROPN_relfreq	-0.352593	0.702863
12	PUNCT_relfreq	-0.602143	0.547637
13	average_sentence_length	-0.744892	0.474785
14	NOUN_relfreq	-0.869392	0.419206
15	AUX_relfreq	-0.937776	0.391498
16	monosyll_relfreq	-1.254602	0.285189
17	word_count	-1.284954	0.276663
18	sentence_count	-1.779108	0.168789
19	stopword_relfreq	-1.933990	0.144570
20	NUM_relfreq	-2.124516	0.119491
21	unique_tokens_relfreq	-3.039943	0.047838

Figure 2, 1st approach to perform model coefficients

For example, the feature ‘CCONJ_relfreq’ has positive log odds of 1.06, which means that coordinating conjunction increases the likelihood of predicting ‘gpt’. On the other hand, the feature ‘unique_tokens_relfreq’ has the highest negative log odds value of -3.04, indicating that unique token frequency increases the likelihood of predicting ‘human’. Both of these features have the highest impact on classification.

Considering that the predictive model may be used by both experts and non-experts, the results and process should be easily interpretable for all users. (Benton) assumed that logistic regression coefficients can be challenging to interpret directly. However, using odds ratios to describe the relationships and relevance of features makes it easier to explain the results, showing how much the odds change. For example, if a feature’s value is greater than 1, it has a positive impact, less than 1 indicates a negative impact and if the feature equals 1, it has no effects and could be excluded. As a model result, the odds ratio of 2.90 for ‘CCONJ_relfreq’ shows that a one-unit increase in coordinating conjunction frequency nearly triples the odds of the text classified as ‘gpt’. In contrast, the odds ratio for ‘unique_tokens_relfreq’ is less

than 0.05, meaning that an increase in unique token frequency leads to the text classified as ‘human’. Furthermore, the model will be adjusted by excluding features like relative frequency of negative words (negative_relfreq), adverbs (ADV_relfreq), determiners (DET_relfreq), subordinating conjunctions (SCONJ_relfreq) or symbols (SYM_relfreq), that have a nearly neutral effect, indicated by odds ratio values close to 1.

The final version of the predictive model includes features from text statistics, part-of-speech (POS) tags, sentiment lexicon, and lexical complexity that are relevant to the model. To analyse the text structure and identify different writing styles, the model counts the number of sentences in the text (sentence_count). It identifies sentence structure, complexity or conciseness by calculating the average number of words per sentence (average_sentence_length). To determine the length of the text, it counts the total number of words (word_count). It assesses the emotional tone of the writing by measuring the proportion of positive words in the text (positive_relfreq). However, negative sentiment words, shown a neutral effect, were excluded from the model. The model also measures the complexity of the writing by calculating the proportion of common words ("the", "is", are, etc.) in relation to the total word count (stopword_relfreq). Additionally, it evaluates the proportion of one-syllable words (monosyll_relfreq) and multi-syllable words (polysyll_relfreq), as well as the number of unique (non-repeated) words compared to the total word count (unique_tokens_relfreq).

To further identification the text's structural style, POS tags are used. These tags help identify the frequency of specific word types, such as coordinating conjunctions (CCONJ_relfreq), adjectives (ADJ_relfreq), proper nouns (PROPN_relfreq), and verbs (VERB_relfreq). By analysing these POS tags, the model can better understand the grammatical structure and writing style.

Model evaluation

The results in figure 3 below demonstrate the effectiveness of the logistic regression model in distinguishing between LLM-generated and human-written text. The confusion matrix, known as error matrix, compares predicted values to actual values. Each row of the matrix represents the actual class, while each column represents the predicted class. The diagonal from the top left to the bottom right of the matrix indicates correct predictions.

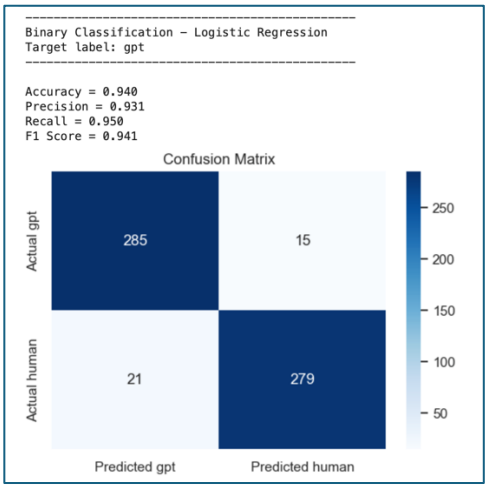

Figure 3, statistical result of logic regression and confusion matrix

Since the base label in this case is 'gpt', the model made the predictions. True Positives, total 285, shows that LLM-generated texts correctly identified as 'gpt'. True Negatives, total 279, shows as human-written texts correctly identified as 'human'. False Positives, as human-written texts incorrectly classified as 'gpt' occurred 15 times while False Negatives, as LLM-generated texts incorrectly classified as 'human' occurred 21 times.

The model correctly predicted 94% of all instances, with an accuracy of 0.940, meaning most texts were classified correctly. The precision value of 0.929 shows that when the model predicted a text as GPT, 92.9% of those predictions were actually correct. The recall value of 0.950 indicates that the model correctly identified 95% of all LLM-generated texts, but it still missed 5% of them. The F1 Score, which balances precision and recall, is high, meaning the model performs well in identifying GPT text while minimising false predictions. The AUROC Score of 0.984 shows the model is very effective at distinguishing between GPT and human texts, with a low chance of classification errors.

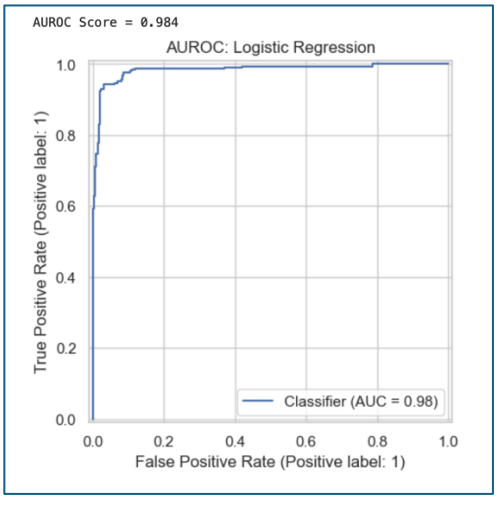

Figure 4, AUROC of logistic regression

The coefficient measures show that none of the features are neutral in this model. Every feature either increases or decreases the likelihood of correctly identifying LLM-generated or human-written text, making them all useful for this task.

Comparison coefficients to statistical summary values of mean, min, max, and standard deviation for each feature in the model gives insight about how typical or extreme values of the features are distributed across the dataset. For example, if the average number of words per sentence (average_sentence_length) has a mean value of around 20.5 words, but the model coefficient shows a negative effect (odds ratio < 1), this suggests that texts with shorter sentences are more likely to be classified as GPT-generated, and longer sentences might lend toward human-written. Features like usage of common words (stopword_relfreq) and number of sentence in the text (sentence_count) have strong negative coefficients. According to the summary, stopword frequency ranges from 0.14 to 0.55, meaning a higher stopword frequency (closer to the max) makes it more likely that the text will be classified as human-written.

```
--------------------------------------------------
Model coefficients
with log odds (logit) converted to odds ratio
for improved interpretability

Target label: gpt
--------------------------------------------------
```

	feature	log odds (logit)	odds ratio
0	CCONJ_relfreq	1.050405	2.858808
1	polysyll_relfreq	0.571353	1.770661
2	ADP_relfreq	0.550877	1.734774
3	PRON_relfreq	0.397507	1.488110
4	VERB_relfreq	0.367446	1.444042
5	positive_relfreq	0.241273	1.272868
6	ADJ_relfreq	-0.228611	0.795638
7	PART_relfreq	-0.317290	0.728119
8	PROPN_relfreq	-0.345390	0.707944
9	PUNCT_relfreq	-0.612875	0.541791
10	average_sentence_length	-0.737798	0.478166
11	NOUN_relfreq	-0.856844	0.424500
12	AUX_relfreq	-0.901833	0.405825
13	monosyll_relfreq	-1.260183	0.283602
14	word_count	-1.351113	0.258952
15	sentence_count	-1.695657	0.183479
16	stopword_relfreq	-1.993775	0.136180
17	NUM_relfreq	-2.135549	0.118180
18	unique_tokens_relfreq	-3.026702	0.048475

Figure 5, final features structure

(Crawford) digs into the hidden political implications of classification systems used in artificial intelligence. She argues that AI systems not just neutral technical systems and critiques them in general illustrating that datasets used for training AI often embed biases related to gender, race, and other social factors.

If the model was trained only on English text, it could struggle with texts in other languages or with non-standard forms of English (e.g., slang, dialects, or jargon). This linguistic bias could cause misclassification when exposed to varied types of texts.

In cases where human-written text is mistakenly classified as GPT, could be problematic, because human text as AI-generated might lead to accusations of plagiarism or academic dishonesty, causing unfair penalties for the author. Therefore, minimising these errors is extremely important.

The model mainly uses basic features like sentence length, word count, or how different parts of speech are used. It doesn't really understand the deeper meaning or intent behind the words. Because of this, it might make mistakes, like incorrectly identifying whether a text is human-written or GPT-generated, especially when it misses the subtle details or context of the text.

The model mainly uses basic features like sentence length, word count, or how different parts of speech are used without understanding of deeper meaning or thoughts behind the words. Because of this, it might make mistakes, incorrectly identify whether a text is human-written or GPT-generated, especially when it misses subtle details or context.

To improve this, comparison between the performance of the logistic regression model with other models like decision trees might be insightful. A decision tree works by breaking down the data into smaller parts and could provide a better understanding of non-linear relationships in the text.

Works Cited

Benton, Jonathan. *Interpreting Coefficients in Linear and Logistic Regression*. 30 June 2020. <https://towardsdatascience.com/interpreting-coefficients-in-linear-and-logistic-regression-6ddf1295f6f1>.

Crawford, Kate. *Microsoft's Kate Crawford: 'AI is neither artificial nor intelligent'*. 2021. <https://www.theguardian.com/technology/2021/jun/06/microsofts-kate-crawford-ai-is-neither-artificial-nor-intelligent>.

Martin, Daniel Jurafsky & James H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models (Third edition)*. 2024.

Rudin, Cynthia. *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. 2019.

Universal Dependencies contributors. *Universal POS tags*. 2024.