Research Article

Fiction vs Non-Fiction Genre Classification: Classical Readability Metrics vs BERT

Rajeshwari Satrasala[1], Kushal Shah[1]

1. Independent researcher

In this paper, we show that fiction vs non-fiction genre classification can be achieved with very high accuracy using simple readability metrics, which have been extensively studied by linguists for many decades. In addition, we explore the BERT model for this classification and find that, although it can also achieve very high accuracy with the same amount of training data, its results are very hard to understand. We tried many adversarial attacks to break the fine-tuned BERT model but found it to be quite resilient.

I. Introduction

The problem of genre identification using linguistically motivated features has been extensively investigated in NLP. However, the particular problem of fiction vs. non-fiction genre classification has started receiving serious attention only in recent years[1][2][3] etc. In[1], this classification was performed based on POS (parts of speech) ratios, and the features implemented by[2] are very elaborate but are not directly connected to cognitive theories, and in[3] the classification of short text/paragraph is analyzed on the ordering of POS. Although these features perform the classification with very high accuracy, they fail to convey a linguistic meaning that is simple for human experts to understand. This work addresses this problem by coming up with meaningful features for fiction and non-fiction genre classification based on age-old readability scores, which have been extensively used in linguistics to quantify the clarity of a given text using the percentage of uncommon or difficult words. Along with the classical Logistic Regression-based classifier, we also fine-tuned the BERT model for this classification to analyze its advantages and disadvantages over the classical method.

 doi.org/10.32388/9A17F4

In English literature, from the early 1880s to the late 1990s, a tremendous amount of research took place on measuring the clarity and accessibility of texts used in classrooms, work environments, and everyday life. This led to the development of readability scores, or simply referred to as readability, which is broadly defined as the ease with which a reader can understand a written text[4]. In this regard, around nine formulas are proposed based on the statistics of the features like average word length, average sentence length, syllable count, uncommon words, or difficult words used, etc. Some of the important formulae in (1b), (1a), and (1c)[5]:

$$\text{SMOG} = \sqrt{polysyllablecount} + 3 \tag{1a}$$

$$\text{Dale Chall} = 0.1579 \times (\text{PWD}) + 0.0496 \times (\text{ASL}) + 3.6365 \tag{1b}$$

$$\text{Spache} = 0.141 \times (\text{ASL}) + 0.086 \times (\text{PWD})_1 + 0.839 \tag{1c}$$

In Eq. (1a), the polysyllable count is the number of words of more than two syllables in a sample of 30 sentences. In Eq. (1b), ASL means Average Sentence Length, PWD is the percentage of difficult words not part of the Dale Chall common word list, and in (1c) PWD_1 is the percentage of difficult words not part of the Spache common word list.

In this paper, the standard Brown corpus[6] and Baby BNC[7] datasets are used to create the corpus for the fiction and non-fiction classification task. The detailed analysis is presented in the following sections.

II. Corpus creation

A. Corpus for Logistic Regression-based classifier

In our analysis, the standard Brown and Baby BNC corpora are used. Out of the 15 genres in the Brown dataset, we excluded the 5 genres of humor, editorial, lore, religion, and letters from our dataset as it is difficult to accurately associate them with either fiction or non-fiction genres. Finally, in the Brown corpus, the fictional category consists of 5 subcategories, namely, fiction, mystery, romance, adventure, and science fiction. Similarly, the non-fiction category includes 5 subcategories, namely, news, hobbies, government, reviews, and learned.

The Baby BNC consists of four categories, namely, fiction, newspaper, spoken, and academic. Due to the clear demarcation between these categories, we use only fiction documents labeled as fiction and academic documents as non-fiction for our experiments.

The Brown corpus has a total of 324 files, out of which the fiction category has 117 files with 1957 words on average, and the non-fiction has 207 files with 1953 words on average. The Baby BNC has a total of 55 files, out of which 25 files belong to the fiction category with an average of 39186 words and 20 files of the non-fiction category with an average of 32560 words.

B. Corpus for BERT-based classifier

For fine-tuning the BERT model, we need to break each file into smaller pieces of text such that each piece of input text has a maximum of 512 tokens (the limit set by the BERT model). Hence, we create the training, testing, and evaluation corpora using the Brown and Baby BNC datasets to have paragraphs of approximately 300 to 350 words with complete sentences such that after tokenization, the token length is within the BERT maximum token length of 512.

III. Classification using Logistic regression

In this section, we present the classification results using a few classical features. In our analysis, the standard Brown and Baby BNC corpora are used. The “py-readability-metrics," a Python package, is used to get the readability scores and grades of different readability formulas, and statistics of the files. The Stanza POS tagger[8] is used for corpus POS tagging. Here, we are considering only the alphabetical words and do not include any white spaces, numerical, or special characters while counting the number of words and the number of difficult words. To ensure this, we have used isalpha() from the NLTK library to filter only the alphabetical words. In our experiments, we have also explored the effect of using min-max scaling on the input data. All the classification results are averaged over 1000 runs with random shuffling and splitting of the Baby BNC and Brown corpus.

In Table I, classification results based on Dale-Chall, Spache, and SMOG readability scores are tabulated. The “Dale-Chall score" alone gives 94% test accuracy on the Brown and 99% test accuracy on the Baby BNC corpus. These results are almost comparable with the two feature-based results proposed in[1].

The Dale-Chall and Spache readability formulas have high similarity, and both are dependent on the uncommon words present in the document/paragraph. These uncommon words are mostly polysyllabic words, and their correlation is tabulated in IV. A large portion of the common words listed in the Dale-Chall[9] list and the common words in Spache[10] overlap with each other.

Features	Testing Accuracy on brown with Brown training	Evaluation accuracy on Baby BNC with Brown training	Testing Accuracy on Baby BNC with Baby BNC training	Evaluation accuracy on brown with Baby BNC training
['smog score']	0.9 ± 0.04	0.99 ± 0.03	0.99 ± 0.04	0.9 ± 0.03
['spache score']	0.95 ± 0.04	0.94 ± 0.03	0.96 ± 0.04	0.87 ± 0.06
['dale chall score']	0.94 ± 0.04	1 ± 0.03	0.99 ± 0.05	0.94 ± 0.03

Table I. Important readability scores based classification results for 80-20 training and test split.

We further analyzed the classification of fiction and nonfiction based on low-level features such as the percentage of difficult words, average sentence length, and the ratio between the number of nouns to the number of verbs and compared their performance with two feature results presented to[1]. The performance of the Logistic Regression-based classifier with these low-level features is tabulated in Table II.

A similar analysis was conducted on the corpus of short texts and paragraphs used as input for the BERT model, with the results summarized in Table III. It was observed that applying isalpha() and MinMax scaling led to an approximate 7% drop in accuracy compared to the values for long paragraphs II. These results align closely with those published in[3], which are based on syntactic and combined features. This indicates that the Ratio of difficult words feature is quite robust to the length of the input paragraph.

Features	Testing Accuracy on brown with Brown training	Testing Accuracy on Baby BNC with Baby BNC training
Percentage of Difficult words (with isalpha and Minmax Scaling)	0.937 ± 0.028	0.967 ± 0.065
Percentage of Difficult words (with isalpha and without Minmax Scaling)	0.76 ± 0.072	0.595 ± 0.22
Percentage of Difficult words (without isalpha and Minmax Scaling)	0.851 ± 0.042	0.862 ± 0.111
Percentage of Difficult words (without isalpha and without Minmax Scaling)	0.661 ± 0.067	0.513 ± 0.185
Ratio of Adjective to Pronoun	0.955 ± 0.022	0.928 ± 0.08
Ratio of Noun to Verb	0.95 ± 0.025	0.999 ± 0.009
Ratio of Adverb to Adjective	0.88 ± 0.036	0.906 ± 0.088
Average number of words per sentence	0.909 ± 0.036	0.949 ± 0.065

Table II. Performance of low-level features for 80-20 train and test split. While using the ratio of difficult words as a feature, it is very important to take only alphabetical words (use the isalpha() function) and apply Minmax() scaling during Logistic Regression training to achieve the best accuracy. For other features, some drop in accuracy is observed with Minmax() scaling. Hence, based on the features, Minmax() scaling is required

doi.org/10.32388/9A17F4

Features	Testing Accuracy on brown with Brown training	Testing Accuracy on Baby BNC with Baby BNC training
Percentage of Difficult words (with isalpha and Minmax Scaling)	0.872 ± 0.014	0.861 ± 0.008
Percentage of Difficult words (with isalpha and without Minmax Scaling)	0.86 ± 0.014	0.82 ± 0.008
Percentage of Difficult words (without isalpha and Minmax Scaling)	0.785 ± 0.016	0.75 ± 0.01
Percentage of Difficult words (without isalpha and without Minmax Scaling)	0.781 ± 0.016	0.74 ± 0.1
Ratio of Adjective to Pronoun	0.653 ± 0.038	0.84 ± 0.09
Ratio of Noun to Verb	0.848 ± 0.014	0.874 ± 0.008
Ratio of Adverb to Adjective	0.75 ± 0.018	0.751 ± 0.01

Table III. The performance of low-level features was analyzed for shorter paragraphs using an 80-20 split between training and testing data. The isalpha() function and MinMax scaling were applied to improve accuracy. While MinMax scaling had a minimal impact on the Ratio of Difficult Words, it led to a drop in accuracy for other features. Therefore, MinMax scaling is necessary based on the specific features being used

Category	corpus	DiffPoly	DiffAdjPron	DiffAdvAdj
Fiction	Brown	0.811	0.4657	-0.503
	Baby BNC	0.907	0.721	-0.770
Non Fiction	Brown	0.838	0.505	-0.643
	Baby BNC	0.847	0.644	-0.756

Table IV. Correlation coefficient between percentage of difficult words, percentage of polysyllable words, ratio of adverb to adjective, ratio of adjective to pronoun, and ratio of noun to verb. DiffPoly = correlation coefficient between percentage of difficult words and percentage of polysyllable words. DiffAdjPron: correlation coefficient between percentage of difficult words and ratio of adjective to pronoun. DiffAdvAdj: correlation coefficient between percentage of difficult words and ratio of adverb to adjective.

In Table IV, we tabulate the correlation coefficients between these low-level features, which shows how these features are related. We can see that there's a strong connection between the percentage of difficult words and the percentage of long (polysyllable) words. Similarly, the percentage of difficult words is strongly linked to the ratio of adjectives to pronouns and the ratio of adverbs to adjectives. When all words in each file are considered for the analysis, the percentage of all difficult words has a very strong negative correlation with the ratio of all adverbs to adjectives. This indicates that the percentage of difficult words set has a high overlap with the ratio of adjectives to pronouns and the ratio of adverbs to adjectives.

doi.org/10.32388/9A17F4

IV. Classification using BERT

Model	Training Data Acc	Test Data Acc
Pre-trained BERT	0.3614	0.3558
Finetuned one epoch BERT	0.9972	0.9908

Table V. Pre-trained original BERT model and fine-tuned on Brown training corpus for one epoch. BERT model accuracy is tabulated.

The BERT base case model is fine-tuned for the classification task using the training corpus. Fine-tuning is done using the corpus created using the Brown dataset with an 80-20 test-train split and one epoch. In Table V, we compare the performance of the pre-trained BERT model with our fine-tuned version. It shows that fine-tuning BERT on the training corpus for one epoch with all other parameters set to their default state is sufficient to achieve 99% accuracy, whereas the pre-trained model only gives around 35% accuracy.

Although the BERT model gives very good accuracy with minimal effort (no need to work hard to figure out the features), its major drawback is that it's a black box and does not provide any insights into how the classification was actually achieved. So, in order to understand what might be going on inside the model, the following experiments were run on the fine-tuned BERT:

1. Without Nouns (WoNu): remove all the nouns from the original test and training corpus.
2. Without Verbs (WoVr): remove all the verbs from the original test and training corpus.
3. Without Pronouns (WoPnu): remove all the pronouns from the original test and training corpus.
4. Without Adverbs (WoAdv): remove all adverbs from the original test and training corpus.
5. Without Adjectives (WoAdj): remove all adjectives from the original test and training corpus.
6. Without Difficult Words (WoDiff): remove all the difficult words (words which are not in the dalle and spache words list) from the original test and training corpus.
7. Without Difficult Words+POS (WODiffxxx): remove all the difficult words and some POS (like noun, pronoun, verb, adjective, adverb) from the original test and training corpus.

8. Word Scrambling (WordSrc): First, break the original paragraphs in the corpus into a list of words and randomly scramble this list of words. Then create a paragraph by joining these scrambled words to replace the original paragraphs in the corpus.
9. Sentence Scrambling (SentSrc): First, break the original paragraphs in the full corpus into sentences and then randomly scramble the sentences (mix the sentences across the files and within the files) and create a paragraph of approximately 300-350 words such that the token length is less than 512.
10. Fixed Length Scrambled Word Sentence (FixSrcWord): First, break the original paragraphs in the corpus into a list of words and randomly scramble this list of words. Then select a fixed number of words from this scrambled word list to create a sentence to replace the original paragraph.

Here are some of our key findings from the above experiments on the BERT model:

1. The classical NLP features used in our Logistic Regression model have none or very limited effect on the BERT behavior, perhaps because BERT sees only the tokens (sub-words) and not any specific kind of words.
2. Scrambling of words or scrambling of sentences has no effect on the accuracy, perhaps because the BERT model does not use rigid positional embeddings.
3. Even as low as 5 random words out of 300 to 350 input words are sufficient for BERT to provide approximately 79% accuracy. As the number of words increases, the accuracy of the BERT model increases.
4. The BERT outputs the same number of embeddings as the number of input tokens. These embeddings represent not only the input tokens but the semantic (joint conditional distribution) of that token with respect to all the other 512 input tokens. Hence, trying to understand what exactly BERT is learning seems to be a hopeless exercise to begin with.

All these observations indicate that BERT is robust to most of the adversarial attacks.

V. Discussion and Conclusion

In this paper, we have shown that fiction vs non-fiction genre classification can be achieved with very high accuracy using the percentage of difficult words in a text. Although this accuracy value is similar to that obtained using the two ratio features (adverb/adjective and adjective/pronoun)[1], an advantage of using the difficult words percentage is that it is easier to relate to. We do expect non-fiction texts to have a higher percentage of difficult words as compared to fiction texts.

We also explored the BERT model and found it to give a very high accuracy of classification. However, the features it may have used for this classification seem very hard to comprehend despite our best efforts. As a person can almost guess the storyline of a movie by viewing the starting and ending movie sciences based his/her prior experience of viewing and remembering many other movies in the past, in a similar way, BERT seem to be able to extract the semantics of an entire paragraphs of 300 words with just 5 to 10 random words from this paragraph. So, the conditional join distribution of words of a paragraph is similar to the prior knowledge about the world that humans use in many of the tasks they perform using their cognitive abilities.

References

1. a, b, c, d, e *Qureshi R, Ranjan S, Rajkumar R, Shah K. (2019). "A simple approach to classify fictional and non-fictional genres". Association for Computational Linguistics, Proceedings of the Second Workshop on Storytelling. :81–89.*
2. a, b *Vicente M, Miró Maestre M, Lloret E, Suárez Cueto A. (2021). "Leveraging machine learning to explain the nature of written genres". IEEE Access. 9:24705–24726.*
3. a, b, c *Kazmi A, Ranjan S, Sharma A, Rajkumar R. (2022). "Linguistically motivated features for classifying shorter text into fiction and non-fiction genre". In: Proceedings of the 29th international conference on computational linguistics. International Committee on Computational Linguistics pp. 922–937.*
4. ^ *Readability. Available from: https://en.wikipedia.org/wiki/Readability*
5. ^ *Readability index in python(NLP). Available from: https://www.geeksforgeeks.org/readability-index-python nlp/*
6. ^ *Francis WN, Kučera H. (1989). "Manual of information to accompany a standard corpus of present-day edited american english: For use with digital computers". Brown University, Department of Linguistics.*
7. ^ *BNCConsortium. (2007). "British national corpus, baby edition". Oxford Text Archive.*
8. ^ *Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. (2020). "Stanza: A python natural language processing toolkit for many human languages". Association for Computational Linguistics (ACL) System Demonstrations.*
9. ^ *Common words acoording to dale chall. Available from: https://help.readable.com/en/article/dale-chall-words-list-w877fe/?_ga=2.128864650.789442227.1731919731-2055000836.1731919731*
10. ^ *Common words acoording to spache. Available from: https://help.readable.com/en/article/what-is-the-spache-words-list-1mxg9cb/*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.