Linguistically Motivated Features for Classifying Shorter Text into Fiction and Non-Fiction Genre

Arman Kazmi[1] **Sidharth Ranjan**[2] **Arpit Sharma**[1] **Rajakrishnan Rajkumar**[1]
IISER Bhopal[1] IIT Delhi[2]
{kazmiarman85, sidharth.ranjan03}@gmail.com
{arpit, rajak}@iiserb.ac.in

Abstract

This work deploys linguistically motivated features to classify paragraph-level text into fiction and non-fiction genre using a logistic regression model and infers lexical and syntactic properties that help distinguish the two genres. Previous works have focused on classifying document-level text into fiction and non-fiction genres, while in this work, we deal with shorter texts which are closer to real-world applications like sentiment analysis of tweets. For the task of short-text classification on the Brown corpus, a model containing linguistically motivated features confers a substantial accuracy jump over a baseline model consisting of simple POS-ratio features found effective in previous work. The efficacy of the above model containing a linguistically motivated feature set also transfers over to another dataset viz, Baby BNC corpus. Subsequently, we compared the classification accuracy of the logistic regression model with two deep-learning models. A 1D-CNN model gives an increase of 2% accuracy over the logistic regression classifier on both datasets. A BERT-based model gives state-of-the art classification accuracies of 97% on Brown corpus and 98% on Baby BNC corpus. Although, both these deep learning models give better results in terms of classification accuracy, the problem of interpreting these models remains an open question. In contrast, regression model coefficients revealed that fiction texts tend to have more character-level diversity and have lower lexical density (quantified using content-function word ratios) compared to non-fiction texts. Moreover, subtle differences in word order exist between the two genres, *i.e.*, in fiction texts *Verbs* precede *Adverbs* in contrast to the opposite pattern in non-fiction texts (inter-alia).

1 Introduction

Written text can be classified into various categories based on its content or writing style. This paper focuses on classifying shorter texts into fiction and non-fiction genres based on their writing style. In general, the *fiction* writing has an imaginative writing style and involves non-factual prose content. In contrast, the *non-fiction* writing deals with actual events, places, and persons and is written purely based on the facts. In some cases, distinguishing between these two writing categories is easy due to the content of the text. However, classification becomes challenging in many instances due to blurry boundaries between them. For example, a short story may contain imaginary characters situated in real-life settings. Therefore it is essential to factor in the writing style of texts while classifying them into fiction and non-fiction genres.

The problem of genre identification using linguistically motivated features has been extensively investigated in NLP (see references in the rest of this section). However, the particular problem of fiction vs. non-fiction genre classification has started receiving serious attention only in recent years (Vicente et al., 2021; Qureshi et al., 2019). In the cited works, different features have been studied for classifying document-level texts (or long texts) into fiction and non-fiction genres. In contrast, very little effort has been expended to investigate the set of relevant features which are effective for the classification of shorter texts, *i.e.*, paragraph-level texts into fiction and non-fiction genres. Shorter texts or paragraph-level texts are more common on the internet and have several important practical applications like breaking news detection, opinion mining, micro-blog summarization, and discovering trending topics (Kateb and Kalita, 2015). Genre identification tools for shorter text can potentially be deployed to filter out specific categories of tweets, news headlines, product reviews, and online app reviews which have been written to manipulate or influence the users/customers. For such applications, fiction vs. non-fiction classification technology capable of analyzing writing styles can play a crucial role.

The main objective of this paper is to identify the

Proceedings of the 29th International Conference on Computational Linguistics, pages 922–937
October 12–17, 2022.

most relevant features that not only enable one to build an effective classifier but also provide deeper insights about the properties that can be used to distinguish these two genres in shorter texts. To this end, we deployed features belonging to four categories, *i.e.*, *Raw* features, *POS-ratio* features, *Lexical* features and *Syntactic* features (Karlgren and Cutting, 1994; Buongiovanni et al., 2019; Biber and Stubbs, 2002; Cleuziou and Poudat, 2007). Raw text features quantify the basic properties of the text, like sentence length and variation in sentence length within a paragraph. Lexical features are based on the statistics of words or characters present in the corpus. In writing, vocabulary plays an important role as it involves the coordination of many higher levels and lower levels of cognitive skills (Hayes, 2000; Olinghouse and Leaird, 2008). Previous studies have also used various measures of lexical diversity to discern differences between genres (Milička and Kubát, 2013; Sadeghi and Dilmaghani, 2013). We deployed a character-level diversity estimate and a lexical density estimate (ratio of content to function words). POS-ratio features proposed by Qureshi et al. (2019) compute the ratios of different parts of speech tags present in the corpus, e.g. ADVERB/NOUN, ADJECTIVE/VERB, and VERB/PRONOUN. They found these ratios to be very effective for document-level fiction vs non-fiction classification (accuracy of 96.31 % on Brown corpus text (Francis and Kučera, 1989)). While most of our features were adapted from prior work, we introduce 3 novel syntactic features extracted from parse trees in this paper. Prominently, we modelled word order variation across genres by extracting *head-dependent* bigrams (containing linear order precedence as well). Inspired from the theoretical psycholinguistics literature, we also incorporated features quantifying *syntactic complexity* (Sampson, 1997; Szmrecsanyi, 2004) as well as *argument-adjunct* patterns (Tutunjian and Boland, 2008).

We extracted the aforementioned features[1] from Brown Corpus paragraphs and performed feature selection experiments using the Recursive Feature Elimination Cross-validation algorithm (RFECV Guyon et al., 2002) on individual and combined feature sets. We report the performance of different classification models trained on several feature combinations and compare them with a baseline model with only two POS-ratio features found effective in prior work (Qureshi et al., 2019). Our classification model containing the best 28 features confers an accuracy score of 91.89% on Brown Corpus (Francis and Kučera, 1989) paragraphs with an accuracy jump of 15.56% over the baseline model containing Qureshi et al.'s simple POS ratio features (76% accuracy on short-text classification in the Brown corpus).

In order to check the transferability and generalizability of our results, we used the aforementioned model trained on the Brown corpus (American English text) to classify shorter texts obtained from the Baby BNC Corpus of British English (Consortium, 2007). Our model obtained an accuracy score of 94% which attests its utility for novel text and demonstrates how it is not biased w.r.t. language variety, *i.e.*, American English (Brown) vs British English (Baby BNC). Following previous work in the NLP literature (Worsham and Kalita, 2018; Kim, 2014; Dauphin et al., 2017), we also compared our classification results based on a traditional logistic regression model (containing handcrafted features) with 2 deep learning models. On shorter text from both Brown and BNC corpora, a 1D CNN model induces a 2% increase in accuracy score over the Logistic Regression classifier. Finally, we used a pre-trained BERT-base-uncased model (Devlin et al., 2018) resulting in state-of-the-art accuracy of 97% on Brown Corpus and 98% on Baby BNC Corpus respectively. Although both the deep learning models (CNN models and the BERT-base-uncased models) result in better results in terms of classification accuracy, they are not easily interpretable *i.e.*, linguistic properties captured by these deep learning models are not obvious.

Another issue is that CNN and BERT models are expensive to train from scratch and are more prone to overfitting when compared to the Logistic Regression classifier. On the other hand, our experimental results using simple logistic regression models are interpretable in terms of the impact of specific features. Our regression coefficients indicate that fiction texts tend to be more diverse in terms of characters and have lower lexical density than non-fiction texts. Subtle differences in word order between the two genres can also be inferred from the coefficients our dependency bigram features. For example, *Verbs* tend to precede *Adverbs* and *Pronouns* in the case of fiction texts, in contrast to the opposite pattern in non-fiction texts.

[1]Scripts to extract various linguistic features used in this work can be accessed here: https://github.com/armankazmi/Linguistic-features-of-text

Genre (#docs)	#Words	#Sentences	#Para
BROWN			
Fiction (207)	63011	4133	764
Non-Fiction (117)	89744	4024	746
BABY BNC			
Fiction (25)	140760	9601	1783
Non-Fiction (30)	34947	1327	243

Table 1: Counts of words, sentences, and paragraphs in Brown and Baby BNC corpora

Our main contribution is that we extend the work of Qureshi et al. (2019) on document-level genre classification to the problem of genre-identification for shorter text by incorporating theoretically and cognitively motivated features. Our best-performing model containing linguistically motivated features substantially outperformed their best-performing model for this novel task. The features deployed by Vicente et al. (2021) (another recent work on fiction vs non-fiction classification cited earlier) are very elaborate but are not directly connected to cognitive theories. Earlier works like Worsham and Kalita (2018) and Mendhakar (2022) analyzed various linguistic characteristics of fictional and non-fictional text but focused more on sub-genre classification within fiction and non-fiction genres.

The rest of the paper is organized as follows. In Section 2 we present details of the data sets used in this study. Section 3 provides the motivation and descriptions of the linguistic features used in this work. Section 4 describes the machine learning algorithms we deployed and the results of genre classification experiments using those algorithms. Section 5 discusses the implications of our findings. Finally, in section 6, we summarize all the results and provide pointers for future research.

2 Data and Methods

Our dataset consists of paragraphs from Brown corpus (Francis and Kučera, 1989) and Baby British National Corpus (Consortium, 2007, BNC). These corpora contain text from fiction and non-fiction genres, thus serve an important resource for our research. We set up a binary classification task to predict shorter texts into fiction and non-fiction genres. Therefore, every long document in these corpora was split into separate paragraphs based on the default paragraph annotation provided. After that, each paragraph was tagged to the class based on the class label of their parent document. To mitigate the data imbalance between the two classes since different paragraphs may have varying lengths in terms of the number of sentences, we chose only those paragraphs that had 5 or 6 sentences, and the rest were discarded. Table 1 provides more details of both the aforementioned datasets.

As a pre-processing step, we automatically tagged and parsed the paragraphs in our dataset using state-of-the-art taggers and parsers. We used Stanza (Qi et al., 2020) for parts-of-speech tagging and Stanford CoreNLP toolkit for dependency and constituency parsing (Manning et al., 2014). The punctuation marks in the paragraphs were stripped off prior to their parsing. We then extracted a wide variety of linguistic features from the tagged and parsed text, thus creating a vector representation of each paragraph. The set of features used for the classification task and the underlying motivation behind using them is described in the subsequent section. We used a traditional machine learning model (logistic regression) as well as two deep learning models (CNN and BERT) for our classification task as described in Section 4.

To further our understanding of our classification models, we tested the model's applicability in British English, where we use British National Corpus (Consortium, 2007). This way, we perform transfer learning where the model is learned on one corpus, and its applicability is tested on another corpus. This also provides a more robust way of analyzing our model's predictions. Baby BNC corpus consists of four categories: *fiction, newspaper, spoken,* and *academic*. Following Qureshi et al. (2019), we considered *academic* documents in non-fiction category and *fiction* documents in fiction category, and rest others were excluded from our primary analyses. As they mention, the news genre lacks a clear demarcation[2] in either category.

3 Linguistic features

For genre classification of shorter texts, we deployed the following four distinct categories of features in our work: 1. Raw text 2. Lexical 3. POS ratios 4. Syntactic features. These features and their motivation are described below.

3.1 Raw Text Features

Raw text features (Buongiovanni et al., 2019) are the most basic features. Following the cited work,

[2]We additionally investigate the *news* category of this corpus and report results in Appendix E to motivate future research direction.

Feature category	Feature sets (#features after RFECV)	Testing Accuracy %	F1 score (fiction: 1)	F1 score (non-fiction: 0)
Baseline	adv/adj, adj/pro	76.33 ± 1.700	0.784 ± 0.013	0.737 ± 0.024
Raw Features	avg_sen_len, std_sen_len	73.36 ± 1.64	0.740 ± 0.019	0.726 ± 0.016
Lexical Features	Character diversity (CD; 4)	81.54 ± 1.637	0.817 ± 0.017	0.813 ± 0.017
	Lexical density (lex_den)	63.89 ± 1.772	0.643 ± 0.017	0.634 ± 0.019
POS Features	POS ratios (4)	81.36 ± 1.143	0.823 ± 0.014	0.802 ± 0.01
Syntactic features	syn_comp (6)	72.98 ± 2.47	0.737 ± 0.025	0.721 ± 0.024
	Argument/Adjunct	78.15 ± 1.335	0.781 ± 0.015	0.782 ± 0.015
	dep_rel (19)	87.64 ± 1.672	0.88 ± 0.017	0.871 ± 0.017
	dep_big (36)	89.65 ± 0.553	0.899 ± 0.006	0.893 ± 0.006
Combined features	CD + POS (7)	87.01 ± 1.208	0.875 ± 0.012	0.864 ± 0.013
	CD + POS + syn_comp (8)	86.55 ± 0.984	0.869 ± 0.011	0.861 ± 0.01
	Best features (28)	91.89 ± 0.883	0.921 ± 0.009	0.916 ± 0.008

Table 2: Classification accuracy using different feature set on Brown corpus paragraphs (random baseline: $49.32 \pm 1.61\%$)

we incorporated the following measures (computed over each paragraph) as features: 1. Average sentence length (*avg_sen_len*) 2. Standard deviation of sentence lengths (*std_sen_len*)

3.2 Lexical Features

Descriptions of the two lexical features used in our approach are given below.

- *Character diversity (CD)* can be measured in various ways by establishing statistical relationships between types and tokens in the text. Generally, words are considered to be the tokens of a text, but in our case, we consider characters (excluding space) in the text as tokens[3]. Diversity establishes the statistical relationship between the type and tokens of the text and has been deployed in various applications, such as measuring the proficiency of a second language learner (Engber, 1995; Karakoç and Köse, 2017), studying the speech of people with mild aphasia (Cunningham and Haley, 2020), and analyzing the writing style of authors.

  The most common approach for measuring the diversity of characters or words is to use the ratio of unique tokens divided by the total number of tokens in a text sample, commonly known as TTR (type-token ratio). One of the shortcomings of TTR-based measures is that they depend on the sample length. Therefore, we have used seven other measures of diversity, *i.e.*, Maas Index (*Maas TTR*) (Mass, 1972), Mean segmental type-token ratio (*MsTTR*), Moving Average type-token ratio (*MATTR*) (Covington and McFall, 2010), Measure of Textual Lexical Diversity (*MTLD*) (McCarthy and Jarvis, 2007), moving average of MTLD (*MTLD MA*), VocD (Durán et al., 2004) and YulesK (Greg and Yule, 1944). The exact mathematical formula for each measure above is provided in Appendix A.

- *Lexical Density (lex_den)* features are calculated by taking the ratio of content words (words that are tagged as noun, verb, adjective adverb) to function words (all part of speech tagged words except those of content words).[4] Lexical density quantifies "how informative a text is". Prior work has argued that a text with a high number of content words carries more information than one with a higher number of function words (Johansson, 2008).

3.3 POS Ratio Features

A total of eight parts of speech ratios (*adverb/noun, adverb/pronoun, adjective/verb, noun/verb, verb/pronoun, adverb/adjective, adjective/pronoun, noun/pronoun*) were extracted from tagged datasets based on their efficacy in document-level genre classification (*i.e.*, fiction vs non-fiction) reported in prior work (Qureshi et al., 2019).

[3]Originally, we considered words as tokens and included it in the lexical feature category. However, our preliminary analysis suggested that character-level tokens performed much better than word-level tokens in our classification task, so we did not include words as tokens in the current work.

[4]Following later works, we deviate from Ure's 1971 original definition of lexical density as the ratio of number to content to all words.

3.4 Syntactic Features

The following measures were used in our analysis.

- *Frequency of dependency relations (dep_rel)*: For each parsed paragraph, we extracted the frequency of dependency relations (as defined in the Universal Dependencies framework: https://universaldependencies.org/

- *Argument-Adjunct Ratio*: Ratio of *arguments* to *adjuncts* in each paragraph. In syntactic theory, an adjunct is an optional component of a sentence, clause, or phrase, while arguments are the obligatory parts of a sentence. In psycholinguistics, the argument-adjunct distinction has been empirically demonstrated to impact parsing *i.e.*, the process of constructing syntactic representations progressively during sentence comprehension (Tutunjian and Boland, 2008).

- *Syntactic complexity (syn_comp)*: In our work, we deployed 3 different indices (capture the complexity of a sentence) proposed in prior work. Sampson (1997) defined a depth measure quantifying the degree of left-branching of a constituency parse tree (*depth*). Sampson verified the claim that English writers tend to avoid grammatical structures where the number of left branches between any word and the root node of a sentence exceeds a specific fixed limit (see Figure 3 in Appendix B for an illustration). Another way to measure the syntactic complexity of a sentence is to calculate the average dependency distance (*add*) in a sentence based on a dependency parse tree (Oya, 2011). The third measure deployed in our study is the index of Syntactic Complexity (*isc*), which is based on counts of linguistic tokens that reflect the degree of embeddings or grammatical properties of the text, such as subordinating conjunctions, Wh-Pronouns, Verb forms and Noun phrases (Szmrecsanyi, 2004). For each of the above measures, we calculated the average complexity and the standard deviation on each paragraph.

- *Dependency bigrams (dep_big)*: For each dependency parse tree corresponding to the sentences in our dataset, we extracted bigrams consisting of the POS tags of each syntactic head and dependent pair in the sentence. The linear order of each head-dependent pair was encoded via the keywords *i.e.*, *before* or *after* (see Figure 4 in Appendix C for an illustration). The main objective of this feature was to model word order variation in the text.

4 Experiments and Results

This section presents the results of our experiments for classifying shorter texts into fiction and non-fiction genre. The following subsections describe our classification results using a traditional machine learning model (logistic regression) and two deep learning models (CNN and BERT).

4.1 Logistic Regression Model

We used LOGISTIC REGRESSION (McCullagh and Nelder, 2019) as one of our classification models for the classification task. We evaluate model performance using classification accuracy and F1 score. We selected the optimal features by applying recursive feature elimination with cross-validation (RFECV) on the 4 feature sets described in the previous section. RFECV discards features from a model by fitting the model several times, removing the weakest-performing feature at each step. After obtaining the optimal features, we trained a logistic regression model with 10-fold cross-validation and L1 regularization using scikit-learn toolkit (Pedregosa et al., 2018) on the following two datasets:

1. Brown corpus with a $70\% - 30\%$ train-test split (Training paragraphs: 1057, Testing paragraphs: 453).

2. Training on Brown corpus and testing on Baby BNC corpus (Training paragraphs: 1510, Testing paragraphs: 10 different sets of 493 paragraphs).

For the first case above, we reported the mean testing accuracy with standard deviation for 10 different combinations of paragraphs in the Brown corpus. And for the second case, we trained the model on the feature vectors of 1510 paragraphs from the Brown corpus and tested it on the Baby BNC corpus. However, as presented in Table 1, the number of fiction paragraphs (1,783) in the Baby BNC corpus exceeds the number of paragraphs in the non-fiction category (243). Therefore, we randomly sampled 250 fiction paragraphs 10 times and combined each set with the 243 non-fiction paragraphs. This approach allowed us to report the

Feature set	Testing accuracy	F1 score (fiction)	F1 score (non-fiction)
Best features (28)	94.016 ± 1.03	0.939 ± 0.0112	0.941 ± 0.009
Baseline (adv/adj, adj/pro)	83.448 ± 1.12	0.843 ± 0.0123	0.824 ± 0.0098

Table 3: Classification accuracy on Baby BNC corpus trained on Brown corpus (random baseline: 50.71%)

Model	Data set	Testing Accuracy (%)	F1 score (fiction)	F1 score (non-fiction)
CNN	Brown Corpus	93.66 ± 0.808	0.939 ± 0.008	0.933 ± 0.008
	Baby BNC Corpus	96.94 ± 0.410	0.968 ± 0.004	0.969 ± 0.003
BERT (base-uncased)	Brown Corpus	97.3 ± 0.64	0.973 ± 0.006	0.972 ± 0.006
	Baby BNC Corpus	98.13 ± 0.486	0.981 ± 0.005	0.981 ± 0.005

Table 4: Classification accuracy of 1D CNN and BERT-base-uncased model on Brown and Baby BNC Corpus

mean testing accuracy with standard deviation on 10 different combinations of fiction and non-fiction paragraphs in Baby BNC corpus. The accuracy of the feature sets was compared with the baseline model containing only two features: adverb to adjective ratio (adv/adj) and adjective to pronoun ratio (adj/pro). These two features were found to be optimal for classifying document-level texts into fiction and non-fiction genre (Qureshi et al., 2019).

4.1.1 Brown Corpus

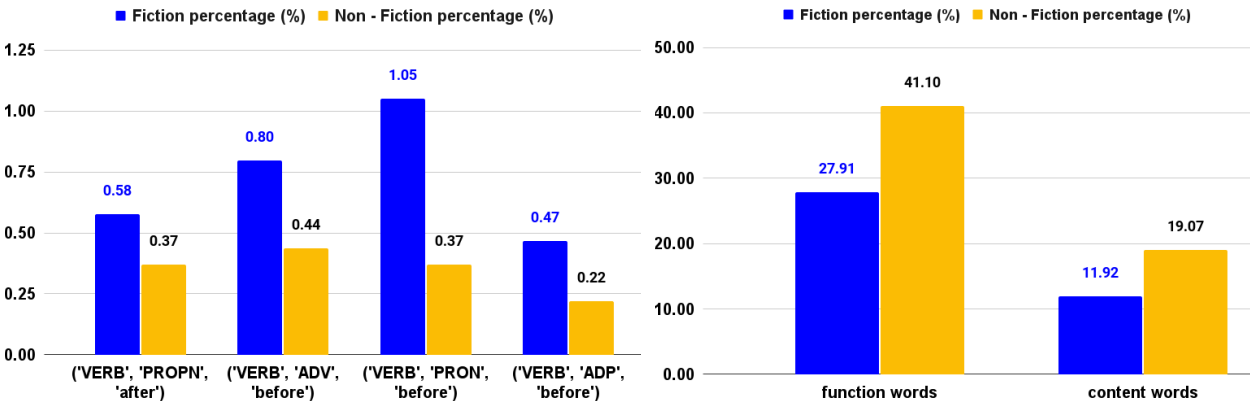
The results of our experiments on the Brown corpus are displayed in Table 2. Individually, the baseline model containing two POS ratio features (adv/adj and adj/pro) gave a classification accuracy of 76.33%. The character diversity (CD) features provided an accuracy gain of 5.21% over the baseline model. However, when CD features are combined with POS-ratio features (CD+POS), the accuracy gain increases to 10.68% over baseline. The syntactic complexity features (syn_comp) performed the worst (72.98%) compared to the baseline. The accuracy significantly improved (86.55%) when syntactic complexity features were combined with CD features and POS-ratio features (CD+POS+syn_comp). The dependency relation distribution features (dep_rel) category returned an accuracy gain of 11.31%. The dependency bigram feature (dep_big) in the syntactic feature category (89.65%) outperformed all other individual feature categories, thus suggesting the significance of our proposed word-order features in this work.

The best performing model contained 28 features after selecting the optimal features from each category after RFECV and gave an overwhelming gain of 15.56% in classification accuracy over the baseline model. Overall, our best-performing model gave a classification accuracy of 91.89%, and F1 scores for each class were similar. Table 5 in Appendix D lists all the optimal features and their regression coefficients that led to the best prediction performance. Interestingly, the feature selection algorithm eliminated both the syntactic complexity (*syn_comp*) features and argument/adjunct ratio features in the syntactic feature category. It is conceivable that dependency bigram features (*dep_big*) would be modeling those generalizations.

We also interpret the coefficients of each predictor in the best-performing regression model to understand their importance for fiction writing. The CD features have positive regression coefficients suggesting that fiction paragraphs tend to be more diverse in terms of characters than non-fiction genres. The negative regression coefficient for the lexical density feature (*content/function* ratio) indicates that fiction paragraphs tend to have lower lexical density than non-fiction paragraphs. In the case of *dep_big* features, 11 features were retained in the optimal feature set. Their coefficients suggest that the fiction paragraphs tend to have more *verbs* after *proper nouns (PROPN)* and more *verbs* preceding *adverbs*, *pronouns* and *Adpositions (ADP)*. In contrast, non-fiction paragraphs tend to have more *proper nouns (PROPN)* before *Numbers (NUM)*.

4.1.2 Baby BNC Corpus

The results of our experiments on the Baby BNC corpus are displayed in Table 3. The classifier yielded a prediction accuracy of 94.01% on the Baby BNC corpus using 28 optimal features obtained previously when the model was trained on the entire Brown corpus. The accuracy score obtained in this case is better than that of the Brown corpus data set, demonstrating that our pre-trained



(a) Dependency bigram features (b) Lexical density features

Figure 1: Genre-wise distribution of features in the Brown corpus

model can be used for a new but related task, *i.e.,* transfer learning. It also suggests that our model is not biased w.r.t. language variety: American English vs. British English.

4.2 Deep Learning Experiments

Recent work has shown the efficacy of neural network-based language models, viz., RNN and LSTMs, for text classification (Bengio et al., 2000; Mikolov et al., 2010; Hochreiter and Schmidhuber, 1997) over traditional n-gram language models (Shannon, 1948, 1951; Chen and Goodman, 1999; Kneser and Ney, 1995; Markov, 1913). While the later models (traditional LMs) struggle with data sparsity and long-range dependencies, the former models (neural net LMs) grapple with substantial memory requirements and a long training time as they work sequentially to capture long-range dependencies (Worsham and Kalita, 2018). The former models also suffer from interpreting the various features learned during their training. In this section, we describe the deep learning experiments performed using CNN (LeCun et al., 1998) and BERT (Devlin et al., 2018) models.

4.3 CNN

We deployed a CNN-based architecture for genre classification, which is inspired by the recent work in the NLP literature (Pham et al., 2016; Prakhya et al., 2017; Dauphin et al., 2017). Recent studies have made use of the CNN-based architecture for tackling some of the challenging NLP problems, including text classification (Kim, 2014; Pham et al., 2016; Dauphin et al., 2017) and learning the abstract linguistic properties of the text, such as inflection, morphological richness, linguistic structure, and word sequence patterns (Prakhya et al., 2017; Rahman et al., 2021). Pham et al. (2016) showed that CNNs are effective in learning language representations up to the sequence of 16 words before the target and can potentially detect high-level abstract features in language data. For the task of genre classification, Worsham and Kalita (2018) compared the efficacy of various deep learning models, including CNN-Kim (Kim, 2014), LSTM (Hochreiter and Schmidhuber, 1997), Hierarchical Attention Network (Yang et al., 2016, HAN) and reported that CNN gave the most reliable performance amidst LSTM and HAN-based deep learning models.[5]

Our CNN experiments involved creating the word embedding vectors using pre-trained GloVe Embeddings. We deployed a 1D CNN model over the embedding vectors to capture the style and patterns in the paragraphs (see Appendix F for more details on training procedures). Table 4 (top block) shows the results of the CNN-based models on Brown and Baby BNC corpora. This model obtains an accuracy score of 93.66% on Brown corpus and 96.94% on Baby BNC corpus.

4.4 BERT

We deployed the BERT-base-uncased model (Devlin et al., 2018) for our genre classification task on shorter texts. The bidirectional encoder representations from transformers (BERT) is an NLP model designed to capture bidirectional represen-

[5]Interestingly, Worsham and Kalita (2018) showed that the XGBoost classifier (Chen and Guestrin, 2016) outperformed every other model deployed for their genre classification task. XGBoost model is based on a tree-based classification algorithm with bag-of-words (BOW) input representation and is known to take the least time and utilize fewer resources.

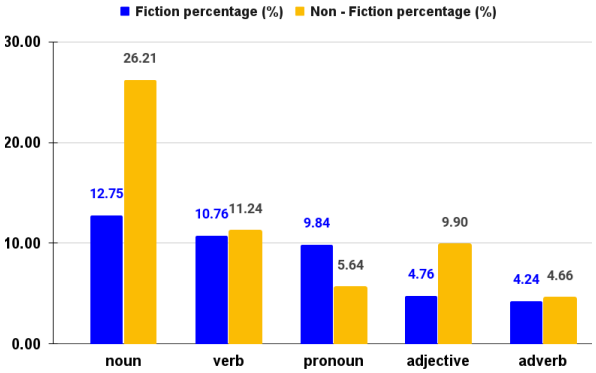


Figure 2: Genre-wise distribution of parts-of-speech (POS) tags in Brown corpus

tation from the unlabeled raw text. Then they are fine-tuned on labeled textual data to carry out various NLP tasks (Vaswani et al., 2017). Appendix G provides more details on the BERT training procedures for our classification task. Table 4 (bottom block) shows the results of the BERT model on Brown and Baby BNC corpora. This model obtains an accuracy score of 97.0% on Brown corpus and 98.13% on Baby BNC corpus.

5 Discussion

Overall, our classification results show that deep learning models achieve better accuracy than the traditional machine learning model for the task of genre identification in shorter text. However, one of the main objectives of this work is to identify the properties of the text that help distinguish between fiction and non-fiction genres. This objective is accomplished using the traditional logistic regression model as the regression coefficients enable interpretation of the features used in the model. Though the deep learning models confer better accuracy on our task, it is difficult to interpret them. The CNN and BERT models used in our work were trained on thousands of parameters. Word embedding vectors capture the properties of the words in the texts, but the complex structure of the internal representations of these models make it difficult to discern the exact generalizations learned by the models in question.

In order to understand the impact of various features in distinguishing between fiction and non-fiction genres, we interpret the regression coefficients of our logistic regression model (depicted in Table 2), by examining the model containing 28 best features (Table 5 of Appendix D). The positive regression coefficient associated with the character diversity feature indicates that fictional text tends to be more diverse in terms of characters than non-fiction text. However, fiction paragraphs have a lower lexical density (negative coefficient of *content/function* ratio) compared to non-fiction ones. This last finding has implications for theories of language production and comprehension, as suggested by a close reading of prior work on sentence processing. Schmauder et al. (2000) showed that during silent reading, both content and function words are processed similarly during the early stages of lexical processing and differently in the latter stages, where words are integrated with other elements of the text (including discourse representations). However, in spontaneous speech, Bell et al. (2009) showed that backward and forward bigram probabilities displayed asymmetric behavior in predicting content and function words, leading to the conclusion that these word types are accessed differently in production.

Dependency relations also provide important cues that help distinguish between fiction and non-fiction genres. Fiction is characterized by a greater number of syntactic subjects, oblique noun modifiers, *'s* possessives, and discourse markers compared to non-fiction. In contrast, non-fiction texts are characterized by a greater frequency of numeral modifiers, passive voice sentences, relative clauses, and multi-word expressions. Further, subtle word order differences between the two genres act as effective predictors of paragraph-level genres. In fictional paragraphs, *verbs* tend to precede *adverbs* and *pronouns* while *proper nouns* are likely to occur before *verbs*. In non-fiction paragraphs, *proper nouns (PROPN)* and *verbs* tend to precede *numbers (NUM)* and *pronouns* precede *verbs*. See Figure 1(a) for a visual illustration of the above patterns in the Brown corpus genres. Similarly, Figure 1(b) depicts the genre-wise percentage of content and function words computed over all the words in the Brown corpus. It suggests that the content and function word percentages in the non-fiction genre are greater than in the fiction genre. Further, Figure 2 represents the percentage of genre-specific parts of speech tags (computed over the total number of Brown corpus parts of speech tags), where the percentage of nouns in fiction is greater than that in non-fiction while *adverbs* and *verbs* have similar distributions across both genres.

In the rest of this section, we illustrate the impor-

tance of our features using linguistic examples (the verb is shown in bold, and the adverb is in italics). Given below are the first two sentences taken from the Brown corpus fiction paragraphs (fileid: cn13, paragraph number: 22):

- The snake **slid** *slowly* and with great care from the new ridge the plow had made , into the furrow and did not **go** any *further*.
- He was multi-colored and graceful and he lay in the furrow and **moved** his arched and tapered head only so *slightly*.

These sentences depict the case where *verbs* precede *adverbs* in fictional text. These examples also indicate how the later adverb plays a crucial role in providing extra information about the verb, thus augmenting the imaginative quotient of the text. We prove two further examples below from non-fiction texts taken from Brown Corpus (fileid: ce16, paragraph number: 17), which shows the genre-specific tendency of *adverbs* preceding *verbs*:

- I laid three layers of glass cloth on the inside of the stem, *also* **installing** a bow eye at this time.
- *Again*, these blocks were **set** in resin-saturated glass cloth and nailed .

Finally, we also checked the performance of all the models (traditional as well as deep learning models) on newswire text from the Baby BNC Corpus (605 paragraphs from the 97 News category documents) individually as well as combined with our Baby BNC dataset (described in Section 2 used in previous experiments). Table 6 in the Appendix section E shows the performance of different models on the Baby BNC corpus when the news texts are included in the non-fiction category. Table 7 of Appendix E shows the percentage of news texts classified as non-fiction using different models. Our traditional model has 68% accuracy while classifying news texts into the non-fiction genre. Even the top-performing BERT-base-uncased model gives a classification accuracy of 83%. The performance drop of these models signifies that news texts contain writing styles that require more detailed linguistic analyses as features found effective for other sub-genres (such as academic texts) fail to achieve a comparable accuracy in this case.

6 Conclusions and Future Work

In this work, we classified paragraph-level text into fiction and non-fiction genres using a traditional machine learning model (logistic regression) and two different deep learning models. For short-text genre identification, we show that the traditional model containing hand-crafted features (raw text, POS ratios, lexical and syntactic features) significantly outperformed a baseline model containing POS-ratio features, (originally proposed by Qureshi et al. (2019) for the task of document-level genre classification). We also obtained the insight that subtle differences in word order exist between the two genres, *i.e.*, in fiction texts *Verbs* precede *Adverbs* (inter-alia) compared to non-fiction texts. Finally, we showed that deep learning models (*viz.*, CNN and BERT) perform significantly better than our traditional model. We obtained state-of-the-art results for the task of short-text genre identification using a pre-trained BERT model fine-tuned on the Brown Corpus.

In future work, we intend to investigate the efficacy of the hand-crafted features on a larger data set and also plan to create a gold standard corpus of human-annotated fiction and non-fiction paragraphs for fine-grained evaluation. Future research needs to investigate whether *syntactic complexity* and *arguments/adjuncts* patterns (not having any impact using our current machine learning setup) are effective predictors of genre shorter texts using other learning algorithms. Further, our finding that journalistic prose (as in news) is not purely non-fiction in nature and might contain fictional elements, needs more systematic investigation. Another line of future inquiry is to combine traditional models (encoding linguistic features) with state-of-the-art deep learning models. Finally, it would be interesting to investigate if causality expressed in the natural language text plays an essential role in classifying text into fiction and non-fiction genres.

Acknowledgements

We would like to thank the COLING-2022 anonymous reviewers, Rupesh Pandey, Kushal Shah, Rameez Qureshi, and Arman Kazmi's MS thesis evaluation committee, for their invaluable comments and feedback on this work. The fourth author also wishes to acknowledge extramural funding from the Cognitive Science Research Initiative, Department of Science and Technology, Government of India (grant no. DST/CSRI/2018/263).

References

Alan Bell, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Douglas Biber and Michael Stubbs. 2002. Dimensions of register variation: A cross-linguistic comparison douglas biber. *Language*, 75.

Chiara Buongiovanni, Francesco Gracci, Dominique Brunato, and Felice Dell'Orletta. 2019. Lost in text. a cross-genre analysis of linguistic phenomena within text. *Italian Journal of Computational Linguistics*.

Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Guillaume Cleuziou and Céline Poudat. 2007. On the impact of lexical and linguistic features in genre- and domain-based categorization. In *CICLing*.

BNC Consortium. 2007. British national corpus, baby edition. Oxford Text Archive.

Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17:100 – 94.

Kevin T Cunningham and Katarina L Haley. 2020. Measuring lexical diversity for discourse analysis in aphasia: Moving-average type–token ratio and word information measure. *Journal of Speech, Language, and Hearing Research*, 63(3):710–721.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Pilar Durán, David Malvern, Brian Richards, and Ngoni Chipere. 2004. Developmental trends in lexical diversity. *Applied Linguistics*, 25.

Cheryl A Engber. 1995. The relationship of lexical proficiency to the quality of esl compositions. *Journal of second language writing*, 4(2):139–155.

W.N. Francis and H. Kučera. 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.

Walter Wilson Greg and G. Udny Yule. 1944. The statistical study of literary vocabulary. *Modern Language Review*, 39:291.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422.

John R Hayes. 2000. Understanding cognition and affect in writing. *Perspectives on writing: Research, theory, and practice*, pages 6–44.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Victoria Johansson. 2008. Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*, 53:61–79.

Dilek Karakoç and Gül Durmuşoğlu Köse. 2017. The impact of vocabulary knowledge on reading, writing and proficiency scores of efl learners. *Journal of Language and Linguistic Studies*, 13:352–378.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Faris Kateb and Jugal Kalita. 2015. Article: Classifying short text in social media: Twitter as case study. *International Journal of Computer Applications*, 111(9):1–12. Full text available.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing*, volume 1, pages 181–184. IEEE.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

A. A. Markov. 1913. Essai d'une recherche statistique sur le texte du roman "Eugene Onegin" illustrant la liaison des epreuve en chain ('Example of a statistical investigation of the text of "Eugene Onegin" illustrating the dependence between samples in chain'). *Izvistia Imperatorskoi Akademii Nauk (Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg)*, 7:153–162. English translation by Morris Halle, 1956.

Heinz-Dieter Mass. 1972. Über den zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 2(8):73.

Philip M. McCarthy and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing*, 24:459 – 488.

Peter McCullagh and John A Nelder. 2019. *Generalized linear models*. Routledge.

Akshay Mendhakar. 2022. Linguistic profiling of text genres: An exploration of fictional vs. non-fictional texts. *Information*, 13(8).

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari.

Jiří Milička and Miroslav Kubát. 2013. Vocabulary richness measure in genres. *Journal of Quantitative Linguistics*, 20:339–349.

Natalie Olinghouse and Jacqueline Leaird. 2008. The relationship between measures of vocabulary and narrative writing quality in second- and fourth-grade students. *Reading and Writing*, 22:545–565.

Masanori Oya. 2011. Syntactic dependency distance as sentence complexity measure. *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2018. Scikit-learn: Machine learning in python.

Ngoc-Quan Pham, German Kruszewski, and Gemma Boleda. 2016. Convolutional neural network language models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1153–1162.

Sridhama Prakhya, Vinodini Venkataram, and Jugal Kalita. 2017. Open set text classification using CNNs. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 466–475, Kolkata, India. NLP Association of India.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Mohammed Rameez Qureshi, Sidharth Ranjan, Rajakrishnan Rajkumar, and Kushal Shah. 2019. A simple approach to classify fictional and non-fictional genres. In *Proceedings of the Second Workshop on Storytelling*, pages 81–89, Florence, Italy. Association for Computational Linguistics.

Chowdhury Rafeed Rahman, MD Rahman, Mohammad Rafsan, Samiha Zakir, Mohammed Eunus Ali, and Rafsanjani Muhammod. 2021. Revisiting cnn for highly inflected bengali and hindi language modeling. *arXiv preprint arXiv:2110.13032*.

Karim Sadeghi and Sholeh Dilmaghani. 2013. The relationship between lexical diversity and genre in iranian efl learners' writings. *Journal of Language Teaching and Research*, 4.

Geoffrey Sampson. 1997. Depth in english grammar. *Journal of Linguistics*, 33:131–151.

A. René Schmauder, Robin K. Morris, and David V. Poynor. 2000. Lexical processing and text integration of function and content words: Evidence from priming and eye fixations. *Memory & Cognition*, 28(7):1098–1108.

Claude E Shannon. 1951. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Benedikt Szmrecsanyi. 2004. On operationalizing syntactic complexity. In *Le poids des mots. Proceedings of the 7th international conference on textual data statistical analysis. Louvain-la-Neuve*, volume 2, pages 1032–1039.

Damon Tutunjian and Julie Boland. 2008. Do we need a distinction between arguments and adjuncts? evidence from psycholinguistic studies of comprehension. *Language and Linguistics Compass*, 2:631–646.

Jean Ure. 1971. Lexical density and register differentiation. *Applications of linguistics*, 23(7):443–452.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Marta Vicente, María Miró Maestre, Elena Lloret, and Armando Suárez Cueto. 2021. Leveraging machine learning to explain the nature of written genres. *IEEE Access*, 9:24705–24726.

Joseph Worsham and Jugal Kalita. 2018. Genre identification and the compositional effect of genre in literature. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1963–1973, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Appendix

A Mathematical formulae for lexical diversity

- **Maas Index:** This measure minimizes the length dependence of TTR by linearizing. Conceptually, this method is based on the notion that a logarithmic curve can reasonably fit the TTR curve (Mass, 1972; McCarthy and Jarvis, 2007).

$$a^2 = \frac{(logTokens - logTypes)}{log^2Tokens} \tag{1}$$

- **MSTTR:** The mean segmental type-token ratio is a metric that divides a text into equal segments based on the amount of words in each segment (normally 50 or 100 words per segment). The TTR is determined for each segment, and the MSTTR is generated by taking the arithmetic mean of the TTR for each segment.

- **MATTR:** The moving average type-token ratio is a measure that involves moving a fixed-size window through the text and calculating the type-token ratio for each window position. To begin, a window length—for example, 50 words—is chosen, and the type-token ratio for words 1–50 is calculated. The type-token ratio is then computed for words 2–51, 3–52, and so on until the text length is reached. The estimated TTRs are averaged for the final score (Covington and McFall, 2010).

- **MTLD:** The measure of textual lexical diversity is defined as the average number of words in a row for which a specified type-token ratio is maintained (here 0.720). When the value falls below a cut-off score (here 0.720), a count (called the factor count) increases by one, and the TTR assessments are reset. It picks up where the value was dropped and repeats the operation until the text is finished. The entire number of words in the text is then divided by the total number of factors in the text. After that, the entire text in the language sample is reversed, and a new MTLD score is calculated. The forward and the reversed MTLD scores are averaged to provide the final MTLD estimate (McCarthy and Jarvis, 2007). One more measure of MTLD was also calculated: Moving Average MTLD (procedure same as that of MATTR).

- **Voc-D:** The vocabulary diversity is a result of a series of random text samplings. It measures the rate at which TTR drops in the sample. To calculate Voc-D, 35 tokens are randomly selected from the text without being replaced, and the TTR is calculated. The average TTR for 35 tokens is estimated and this method is repeated 100 times. Similarly, the average TTR for 36-50 tokens is determined. The means of each of these samples are then used to generate an empirical TTR curve. Using the least-squares approach, a theoretical curve is created that maximizes its fit to the empirical TTR curve. The TTR calculated using Voc-D, 'D' is as follows (Durán et al., 2004).

$$TTR = \frac{D}{N}[(1 + 2\frac{N}{D})^{\frac{1}{2}} - 1] \tag{2}$$

- **Yule's K:** This measures the repetition and lower values of Yule's K represent higher diversities (Greg and Yule, 1944). The value K for a text sample is calculated as follows.

$$K = 10^4 \frac{\{\sum_{r=1}^{N} V_r r^2\} - N}{N^2} \tag{3}$$

where V_r is the number of types that occur r times in a text of length N.

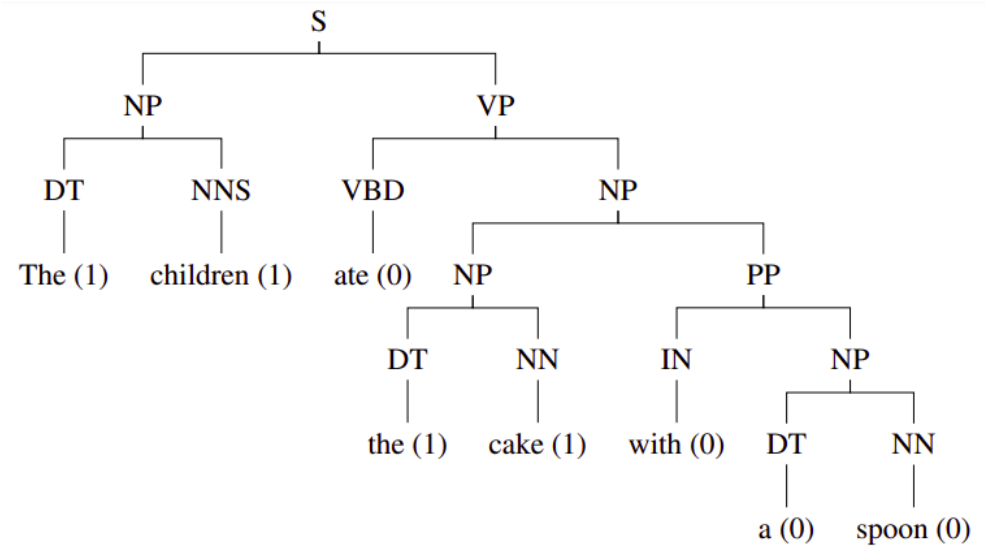


Figure 3: Constituency parse tree indicating Sampson's depth for each word alongside in bracket

B Syntactic depth-based feature calculation

For the example sentence, **"The children ate the cake with a spoon"**, we describe the method to compute Sampson's (1997) measure of syntactic complexity. Figure 3 illustrates the constituency parse tree of the example sentence. Now, we define the LINEAGE of a word as the class of nodes, including the leaf node (terminal node) associated with that word, the root node of its tree, and all the intermediate nodes on the unique path between leaf and root nodes. Now, according to Sampson depth of a terminal node is defined as the total number of those non-terminal nodes in the word's lineage with at least one younger sister.[6] The depth of the word 'cake' in the example sentence according to the above definition is 1. Similarly, the depth of each terminal node or the word in the sentence in sequence from left to right is 1,1,0,1,1,0,0,0, which sums to 4. The depth-based measure is the average over the leaf nodes; hence, the value is 0.5 in this example.

C Dependency bigram feature calculation

For the example sentence, **"The children ate the cake with a spoon"**, we describe the method to compute the dependency bigram feature modelling word order patterns. We extracted bigram features from dependency trees (exemplified in Figure 4). We took all the pos tags of head and dependent pairs from the tree, and specified the position of syntactic heads w.r.t to their dependents in a linear string by means a keyword: *before* or *after*. Therefore, the dependency bigram features for the example sentence are: *(NOUN, DET, after), (VERB, NOUN, after), (VERB, NOUN, before), (VERB, NOUN, before), (NOUN, DET, after), (NOUN, SCONJ, after), (NOUN, DET, after)*

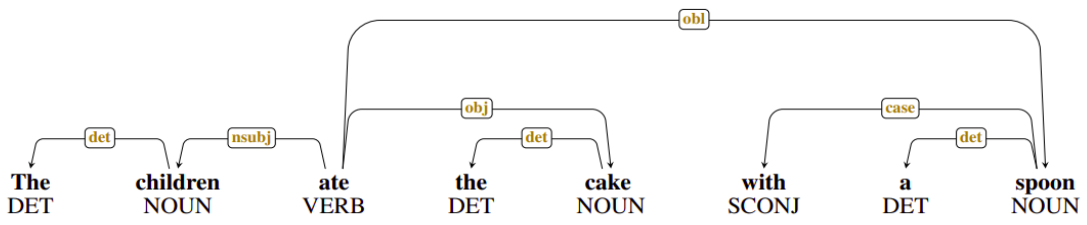


Figure 4: Dependency parse tree to compute the word order based feature.

[6]Node e is a YOUNGER SISTER of a node d if d and e are immediately dominated by the same mother node and e is further right than d.

D Supplementary Information: Brown corpus results

Feature category	Feature set	Feature Name	Regression Coefficient
Raw features	-	std_sen_len	0.11
Lexical features	Character diversity (CD)	TTR	2.31
		Maas TTR	1.70
		VocD	0.38
	Lexical density (lex_den)	content/function	-5.83
POS Features	POS ratios	adverb/pronoun	-0.19
		noun/verb	-0.68
Syntactic features	Dependency relation counts (dep_rel)	discourse	1.17
		nsubj	0.43
		obl:npmod	0.20
		nmod:poss	0.19
		nummod	-0.12
		mark	-0.18
		aux:pass	-0.20
		flat	-0.33
		acl:relcl	-0.49
		fixed	-0.70
	Dependency bigrams (dep_big)	('VERB', 'PROPN', 'after')	0.59
		('VERB', 'ADV', 'before')	0.44
		('VERB', 'PRON', 'before')	0.31
		('VERB', 'ADP', 'before')	0.31
		('PROPN', 'PROPN', 'after')	-0.30
		('VERB', 'SCONJ', 'after')	-0.31
		('VERB', 'NUM', 'before')	-0.45
		('PRON', 'NOUN', 'before')	-0.57
		('PRON', 'VERB', 'before')	-0.74
		('ADJ', 'SCONJ', 'after')	-0.79
		('PROPN', 'NUM', 'before')	-1.46

Table 5: Regression coefficients of the features from our best model containing 28 optimal features

E Supplementary Information: Complete Baby BNC corpus (incl newswire text) results

Model	Testing Accuracy (%)	F1 Score (fiction)	F1 score (non-fiction)
traditional model (28 best features)	83.77 +/- 0.291	0.848 +/- 0.003	0.825 +/- 0.003
GloVe Embedding CNN	87.034 +/- 0.467	0.876 +/- 0.005	0.863 +/- 0.004
BERT-base-uncased	92.52 +/- 0.258	0.927 +/- 0.003	0.922 +/- 0.002
Qureshi et al. (2 best ratio features)	72.95 % +/- 0.4	0.765 +/- 0.004	0.681 +/- 0.003

Table 6: Testing accuracy on Baby BNC corpus including the news texts (Non-Fiction: 848 paragraphs; Fiction: 850 paragraphs; most frequent baseline is 50.05%)

Models	Percentage of samples classified as Non-Fiction
Traditional model (28 best features)	68.42
GloVe Embedding CNN	82.97
BERT-base-uncased	83.3
Qureshi et al. baseline (2 best ratio features)	49.1

Table 7: Percentage of news texts in Baby BNC corpus classified as non-fiction; total samples: 605 news paragraphs

F CNN Training Regime

The pre-trained model of glove embeddings were trained on a text dataset consisting of Wikipedia articles and Gigaword-5 data (collection of newswire texts). The pre-trained vectors were trained on a total of 6B tokens and 400K vocabulary with different embedding dimensions. The model outputs an embedding vector of dimension 100 for each word of the tokenized text. However, depending on the length of tokenized text, the output vectors could be of different lengths. Thus could potentially create an imbalance problem while feeding the vectors into a deep learning model as the input text may not be of fixed length. To overcome this issue, we fixed the length of the input to be the longest sequence length available in the training data. As a result, we obtained an input of constant size for training and testing. The embedding dimension received from each paragraph input is 292×100, where 292 is the fixed maximum length of the tokenized paragraph text in the training data, and 100 is the dimension of each word embedding vector. We pass this embedding layer into a 1D CNN model with 100 filters each of size 3 (also known as kernel size). The activation function used with the CNN layer is 'ReLu'. The output of this layer goes to a global max pooling layer that returns the max value of the input vector received. The output of the global pooling goes to a dense layer of size 10 with the 'ReLu' activation function. Finally, predictions are made using a dense layer of size 1 and a sigmoid activation function, which transforms its output to class probability.

G BERT Training Regime

The BERT-base model (Vaswani et al., 2017) contains an encoder with 12 transformer blocks, 12 self-attention heads, and a hidden size of 768. BERT generates a representation of the sequence from an input sequence raning up to 512 tokens. The sequence consists of one or two segments, with the [CLS] token serving as the sequence's first token and containing the special classification embedding. [SEP] serves as the sequence's second token and is used to separate segments. For text classification, BERT takes the final hidden state of the token [CLS], where the entire sequence information is encoded in this particular token. In the last step, a simple softmax classifier is added to the top of BERT to predict the probability of target labels. We use the BERT-base-uncased model (Devlin et al., 2018) having a hidden size of 768, 12 transformer blocks 12 self-attention heads. The maximum length of the BERT model input was fixed to 512. We then fine-tuned the BERT model with a batch size of 10, a learning rate of 2e-5, and a weight decay of 0.01. The hidden dropout probability was 0.1. We set the maximum number of epochs to 3 and saved the best model for evaluation.