A Simple Approach to Classify Fictional and Non-Fictional Genres

Mohammed Rameez Qureshi
IISER Bhopal
mohr@iiserb.ac.in

Sidharth Ranjan
IIT Delhi
sidharth.ranjan03@gmail.com

Rajakrishnan P. Rajkumar
IISER Bhopal
rajak@iiserb.ac.in

Kushal Shah
IISER Bhopal
kushals@iiserb.ac.in

Abstract

In this work, we deploy a logistic regression classifier to ascertain whether a given document belongs to the fiction or non-fiction genre. For genre identification, previous work had proposed three classes of features, *viz.*, low-level (character-level and token counts), high-level (lexical and syntactic information) and derived features (type-token ratio, average word length or average sentence length). Using the Recursive feature elimination with cross-validation (RFECV) algorithm, we perform feature selection experiments on an exhaustive set of nineteen features (belonging to all the classes mentioned above) extracted from Brown corpus text. As a result, two simple features *viz.*, the ratio of the number of adverbs to adjectives and the number of adjectives to pronouns turn out to be the most significant. Subsequently, our classification experiments aimed towards genre identification of documents from the Brown and Baby BNC corpora demonstrate that the performance of a classifier containing just the two aforementioned features is at par with that of a classifier containing the exhaustive feature set.

1 Introduction

Texts written in any human language can be classified in various ways, one of them being fiction and non-fiction genres. These categories/genres can either refer to the actual content of the write-up or the writing style used, and in this paper, we use the latter meaning. We associate *fiction* writings with literary perspectives, i.e., an imaginative form of writing which has its own purpose of communication, whereas *non-fiction* writings are written in a matter-of-fact manner, but the contents may or may not refer to real life incidents (Lee, 2001). The distinction between imaginative and informative prose is very important and can have several practical applications. For example, one could use a software to identify news articles, which are expected to be written in a matter-of-fact manner, but tend to use an imaginative writing style to unfairly influence the reader. Another application for such a software could be for publishing houses which can use it to automatically filter out article/novel submissions that do not meet certain expected aspects of fiction writing style.

The standard approach in solving such text classification problems is to identify a large enough set of relevant features and feed it into a machine learning algorithm. In the genre identification literature, three types of linguistic features have been discussed *i.e.*, *high-level, low-level* and *derived features* (Karlgren and Cutting, 1994; Kessler et al., 1997; Douglas, 1992; Biber, 1995). High-level features include lexical and syntactic information whereas low-level features involve character-level and various types of token count information. The *lexical* features deal with word frequency statistics such as frequency of content words, function words or specific counts of each pronoun, etc. Similarly, the *syntactic* features incorporate statistics of parts of speech, i.e., noun, verb, adjectives, adverbs and grammatical functions such as active and passives voices or affective markers such as modal auxiliary verbs. The *character-level* features involve punctuation usage, word count, word length, sentence length. And, lastly, the *derived* features involve ratio metrics such as type-token ratio, average word length or average sentence length based information. Majorly, all the previous work involved a combination of different features to represent a particular nature of the document and developing a model that classify different genres, sentiments or opinions.

Notably, researchers have adopted the frequentist approach (Sichel, 1975; Zipf, 1932, 1945) and used lexical richness (Tweedie and Baayen, 1998) as a prominent cue for genre classification (Bur-

Proceedings of the Second Storytelling Workshop, pages 81–89
Florence, Italy, August 1, 2019.

rows, 1992; Stamatatos et al., 2000, 1999). These studies vouch that coming out with statistical distribution from word frequencies would be the *de-facto-arbiter* for document classification. In this regard, Stamatatos and colleagues have shown that most frequent words in the training corpus as well as in the entire English language are one of the good features for detecting the genre type (Stamatatos et al., 2000). With respect to syntactic and semantics properties of the text, previous studies have used various parts of speech counts in terms of number of *types* and *tokens* (Rittman et al., 2004; Rittman, 2007; Rittman and Wacholder, 2008; Cao and Fang, 2009). Researchers have tried to investigate the efficacy of counts vs. ratio features and their impact on the classification model performance. In general, a large number of features often tend to overfit the machine learning model performance. Hence, concerning the derived ratio features, Kessler et al. (1997) argues in his genre identification study that ratio features tend to eliminate over-fitting as well as high computational cost during training.

Although these earlier approaches have made very good progress in text classification, and are very powerful from an algorithmic perspective, they do not provide many insights into the linguistic and cognitive aspects of these fiction and non-fiction genres. The main objective of our work is to be able to extract the features that are most relevant to this particular classification problem and can help us in understanding the underlying linguistic properties of these genres. We begin by extracting nineteen linguistically motivated features belonging to various types (described at the outset) from the Brown corpus and then perform feature selection experiments using Recursive feature elimination with cross-validation (RFECV) algorithm (Guyon et al., 2002). Interestingly, we find that a classifier containing just two simple ratio features *viz.*, the ratio of the number of adverbs to adjectives and number of adjectives to pronouns perform as well as a classifier containing an exhaustive set of features from prior work described above [96.31% and 100% classification accuracy for Brown (Francis and Kučera, 1989) and British National corpus (BNC Baby, 2005), respectively]. This is perhaps the best accuracy reported in the literature so far to the best of our knowledge. Essentially, we find that texts from the fiction genre tend to have a higher ratio of adverb to adjectives, and texts from the non-fiction genre tend to have a higher ratio of adjectives to pronouns. We discuss the implications of this finding for style guides for non-fiction writing (Zinsser, 2006) as well as standard advice proffered to creative writers (King, 2001).

Genre	Subgenre	No. of words	No. of files
Non-fiction	Government	70117	30
	News	100554	44
	Learned	181888	80
	Hobbies	82345	36
	Reviews	40704	17
Fiction	Science Fiction	14470	6
	Fiction	68488	29
	Romance	70022	29
	Adventure	69342	29
	Mystery	57169	24

Table 1: Brown corpus subgenre details

In Section 2, we share details about our linguistic features design, data set and experimental methodology. Section 3 presents the experiments conducted as a part of our study and discusses their critical findings. Finally, Section 4 summarizes the conclusions of the study and discusses the implications of our findings.

2 Data and Methods

For our experiments, we use the Brown Corpus (Francis and Kučera, 1989), one of the earliest collections of annotated texts of present-day American English and available free of cost with the NLTK toolkit (Loper and Bird, 2002). The nature of the distribution of texts in the Brown corpus helps us to conduct our studies conveniently. The Brown corpus consists of 500 text samples with different genres distributed among 15 categories/genres, which are further divided into two major classes, namely, *Informative prose* and *Imaginative prose*. As per our proposed definition in this study, we associate informative prose with the non-fictional genre and imaginative prose as a fictional one. We conduct a binary classification task to separate text samples into these two genres (i.e., fiction and non-fiction) with our proposed linguistic features. Out of the 15 genres, we excluded the 5 genres of *humour, editorial, lore, religion* and *letters* from our dataset as it is difficult to accurately associate them with either fiction and non-fiction genres. Finally, 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. This leads us to use 324 samples out of 500 articles in the Brown corpus; out of which 207 samples fall under fiction category and 117 under non-fiction. Despite having less number of samples, the total word count of all texts in the non-fiction category/genre (479,708 words) is higher than that of fiction (284,429 words), and the total number of sentences in the non-fiction category/genre (21,333) is also higher than that of fiction (18,151). Hence, we chose to divide the data by sub-categories rather than having a number of samples or number of words as the base for distribution. Table 1 provides more details regarding the documents in these genres.

To further our understanding of the model's classification performance for Brown corpus and investigate its applicability to British English, we use the British National Corpus (BNC Baby, 2005). This approach helps us to examine model prediction more robustly. 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 (25 samples) labeled as fiction and academic documents (30 samples) as non-fiction for our experiments. Finally, we apply our algorithm on the articles in the news category (97 samples) to check whether they fall under fiction or non-fiction genre.

Keeping in mind the binary nature of our classification task, we use logistic regression (LR) as our numerical method (McCullagh and Nelder, 1989). Among many classification algorithms, the result of LR is among the most informative ones. By informative, we mean that it not only gives a measure of the *relevance* of a feature (coefficient value) but also the *nature* of its association with the outcome (negative or positive). It models the binary dependent variable using a linear combination of one or more predictor values (features) with the help of following equations where ϕ is the estimated response probability:

$$g(\phi) = log(\phi/(1-\phi)) \quad (1)$$

$$\phi = P(x) = \frac{1}{1 + e^{-(x_i\beta_i + \beta_0)}} \quad (2)$$

where, x_i is the feature vector for text i, β_i is the estimated weight vector, and β_0 is intercept of the linear regression equation.

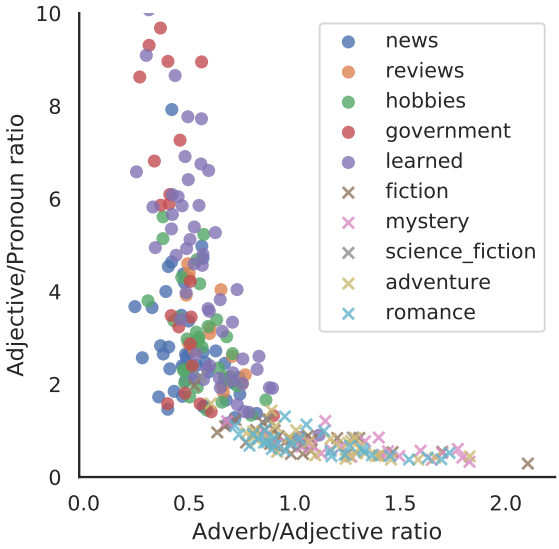

Figure 1: Scatter plot of Brown corpus samples of different subgenres. Fiction samples are marked with 'x' whereas non-fiction samples are marked with 'o' with Y-axis limit set up to 10.

3 Experiments and Results

This section describes our experiments aimed to classify texts into the fictional and non-fictional genres using machine learning. The next subsection describes various linguistic features we deploy in detail and the use of feature selection to identify the most useful features. Subsequently, Section 3.2 provides the results of our classification experiments.

3.1 Linguistic Features and Feature Selection

We compute different low-level and high-level features as discussed in Section 1 and after that take their ratios as the relative representative metric for the classification task. Table 2 depicts the features used in this work. Some of the ratio features such as average token/type (punctuation) ratio, hyphen exclamation ratio, etc., have been explored in earlier work (Kessler et al., 1997). For calculating high-level ratio features, we use tags from two kind of POS tagsets, *i.e*, gold standard tags provided as part of the Brown Corpus (87 tags) and automatic tags (based on the 36-tag Penn Treebank tagset) predicted by Stanford tagger[1] (Toutanova et al., 2003). Grammatical categories like noun, verb, and adjective are inferred from the POS tags using the schema given in Ta-

[1] https://nlp.stanford.edu/software/tagger.shtml

Type	Features
Low-level normalized	**Average Sentence Length** **Average Word Length** Standard Deviation of Sentence Length Standard Deviation of Word Length
Low-level ratio	Average token/type Standard Deviation of token/type Average token/type (punctuation) Standard Deviation of token/type (punctuation) **Hyphen/Quote** Hyphen/Exclamation Quote/Question
High-level ratio	Adverb/Noun Adverb/Pronoun Adjective/Verb Noun/Verb Verb/Pronoun **Adverb/Adjective** **Adjective/Pronoun** **Noun/Pronoun**

Table 2: Derived linguistic features (Features selected after RFECV on: Brown tagset-bold; Penn tagset-underlined)

Category	POS tag	Tagset
Adjective	JJ, JJR, JJS	
Adverb	RB, RBR, RBS, WRB	
Noun	NN, NNS, NNP, NNPS	Penn
Verb	VB, VBD, VBG, VBN,VBP, VBZ	Treebank
Pronoun	PRP, WP	
Adjective	JJ, JJR, JJS, JJT	
Adverb	RB, RBR, WRB, RBT, RN, RP, NR	
Noun	NN, NNS, NN$, NNS$, NP, NP$, NPS, NPS$	Brown
Verb	VB, VBD, VBG, VBN, VBP, VBZ	
Pronoun	PN, PN$, PP$, PP$, PPL, PPLS PPO, PPS, PPSS, PRP, PRP$, WP$, WPO, WPS	

Table 3: Rules to ascertain grammatical categories from POS tags

ble 3. We consider both personal pronouns and *wh*-pronouns as part of the pronoun category.

We use the recursive feature elimination with cross-validation (RFECV) method to eliminate non-significant features. Recursive feature elimination (Guyon et al., 2002) follows the greedy search algorithm to select the best performing features. It forms models iteratively with different combinations of features and removes the worst performing features at each step, thus giving the set of best performing set of features. The motivation behind these experiments is not only to get a good accuracy score but also to decipher the importance of these features and to understand their impact on writing. After applying RFECV on the automatically tagged Brown Corpus, we get all features as the optimum set of features. We attribute this result to the POS-tagging errors introduced by the Stanford tagger. So we apply our feature selection method to features extracted from the Brown Corpus with gold standard tags. Here, 13 out of 19 features are marked as non-significant, and we obtain six most significant features (shown in bold in Table 2). Subsequently, we extract these six features from the automatically tagged Brown Corpus, and feature selection on this set revealed only two of these features as being the most significant (underlined in Table 2). The two most notable features which emerge from our second feature selection experiment are *adverb/adjective ratio* and *adjective/pronoun ratio*. The *Noun/pronoun* ratio feature gets eliminated in the process. Figure 1 illustrates how both these ratios provide distinct clusters of data points belonging to the fiction and non-fiction genres (and even their subgenres). Thus, the Brown corpus tagset encoding finer distinctions in grammatical categories (compared to the Penn Treebank tagset), does help in isolating a set of six significant ratio features. These features are useful for identifying the final two POS-ratios based on automatic tags.

3.2 Classification Experiments

As described in the previous section, we apply logistic regression to individual files of two data-sets (Brown Corpus and Baby British National Corpus) after extracting various low-level features and features encoding ratios of POS tags based on automatic tags emitted by the Stanford tagger (see Table 2). We use a logistic regression classifier with ten-fold cross-validation and L1 regularization for training to carry out our analyses and report the *accuracy* achieved over the total number of files in our test sets. We use the Scikit-learn[2] (Pedregosa et al., 2011) library for our classification experiments. The individual performance by non-significant features has not been reported in our study. We report results for three data sets after tagging them using the Stanford POS-tagger:

1. Brown Corpus with a 60%-40% train-test split (194 training files; 130 test files).
2. Brown Corpus with Baby BNC combined with a 60%-40% train-test split (227 training files; 152 test files).
3. Testing on Baby BNC with Training on Brown Corpus (324 training files; 55 test files).

We calculate testing accuracy for the first two datasets for ten different combinations of training and testing sets, and report the mean accuracy with

[2] https://scikit-learn.org/stable/

S.No.	Data	Feature Sets	Testing Accuracy %	F1 score (non-fiction)	F1 score (fiction)	Baseline Accuracy %	Accuracy Gain %
(a)	Brown Corpus with 60 % - 40 % Train - Test data	All Low level features	94.15 $\pm$ 1.82	0.9540 $\pm$ 0.0141	0.9194 $\pm$ 0.0269	63.77 $\pm$ 2.00	83.85
		19 Features	96.92 $\pm$ 1.26	0.9760 $\pm$ 0.0095	0.9569 $\pm$ 0.0186		91.50
		6 Features	96.08 $\pm$ 1.51	0.9692 $\pm$ 0.0122	0.9457 $\pm$ 0.0206		89.18
		2 Features	96.31 $\pm$ 0.49	0.9711 $\pm$ 0.0038	0.9486 $\pm$ 0.0081		89.82
(b)	Brown Corpus and Baby BNC combined with 60 % - 40 % Train - Test data	All Low level features	95.39 $\pm$ 1.72	0.9634 $\pm$ 0.0138	0.9371 $\pm$ 0.0257	63.13 $\pm$ 3.13	87.50
		19 Features	96.73 $\pm$ 1.73	0.9736 $\pm$ 0.0143	0.9565 $\pm$ 0.0233		91.13
		6 Features	97.21 $\pm$ 1.42	0.9777 $\pm$ 0.0117	0.9624 $\pm$ 0.0196		92.43
		2 Features	97.13 $\pm$ 1.04	0.9769 $\pm$ 0.0087	0.9617 $\pm$ 0.0138		92.22
(c)	Training on Brown Corpus & Testing on Baby BNC	All Low level features	92.73	0.9286	0.9259	54.54	84.01
		19 Features	52.73	0.2353	0.6579		-3.98
		6 Features	100	1	1		100
		2 Features	100	1	1		100

Table 4: Classification accuracy for Brown Corpus and Baby BNC with different feature sets (most frequent class *i.e.*, *non-fiction* baseline results reported).

standard deviation for the same as well as for the most frequent baseline accuracy. While for the third dataset, only one training and testing set possible exists, and therefore, we report the testing accuracy and the most frequent class baseline accuracy accordingly. The most frequent class baseline is the percentage accuracy obtained if a model labels all the data points as the most frequent class in the data (*non-fiction* in our study). Table 4 illustrates our results. Here, we also use another performance metric known as *accuracy gain* which is considered more rigorous and interpretable measure as compared to the standard measure of accuracy. The accuracy gain percentage is calculated as:

$$\texttt{Accuracy Gain \%} = \frac{(acc - baseline)}{(100 - baseline)} \times 100 \quad (3)$$

where *'acc'* is the reported mean accuracy of model, whereas *'baseline'* is the mean of most frequent class baseline.

We begin with the Brown Corpus and take 117 sample texts of non-fiction and 207 of fiction categories. Our training set consists of 60% of the total sample size whereas testing set comprises of remaining 40% of samples. We have four combinations of the set of features (refer Row 1 of Table 4). It can be noted that two features model performed better than the model corresponding to the six features and low-level ratio features and is performing as good as 19 features model. To make the model more robust, we follow the same approach for the combination of Brown corpus and Baby BNC with 147 sample texts of non-fiction and 232 sample texts of fiction categories. Baby BNC has been included to check the impact of British English on the performance of the model. One may observe that the model performed even better when exposed to Baby BNC. Similar observations can be made about the accuracy of the two features model (refer Row 2 of Table 4). In our final experiment, we use the Brown corpus for training and the Baby BNC for testing with the available set of features. In this case, the features obtained after feature selection on the exhaustive set of features results in 100% classification accuracy (Row 3 of Table 4). This result also signifies the universal applicability of the ratio features and high-level POS ratios are not something which is affected by bias due to the language variety (*i.e.*, British vs. American English). However, the low performance of the 19 features model (53% classification accuracy) shows how they are prone to overfitting.

The two most significant features, *adverb/adjective* ratio and *adjective/pronoun* ratio have regression coefficients 2.73 and -2.90 respectively. Thus, fiction documents tend to have higher values for the ratio of number adverbs to adjectives and a lower value for the ratio of the number of adjectives to pronouns. It is worth noting that the high accuracy scores of more than 95% we obtained by using 19 features in the case of the first two datasets are in the vicinity of the accuracy score given by only these two features. Also, the fact that the F1 scores are close to the accuracy values signifies the fact that the results obtained are robust in nature.

Finally, in order to check the dominant tendencies in the behaviour of classifiers containing different feature sets, we examine the predictions of various classifiers using a separate test set consisting of 97 news documents in the Baby BNC corpus. We also studied model predictions using different training sets. Initially, we use the same data sets mentioned in the last two rows of Ta-

Training data	19 features	6 features	2 features
Brown (324 training files)	100	92.78	89.69
Brown (w/o news category) (280 training files)	100	92.78	90.72
Brown + Baby BNC (379 training files)	1.03	92.78	90.72
Brown + Baby BNC (w/ news category) (476 training files)	65.98	97.94	93.81

Table 5: Percentage of documents classified as non-fiction in a test set of 97 Baby BNC news documents

ble 4. Apart from this, to check the bias of the model, we create a new test set after removing the news category from the *non-fiction* class of brown corpus. Similarly, in the combined Brown+Baby BNC corpus, we later include news samples from Baby BNC to measure the improvement in the model's predictions. The results are shown in Table 5. It can be observed that most of the samples are classified as non-fiction, as expected. Also, removing news articles from the Brown corpus non-fiction category does not impact the results indicating the unbiased behavior of the model. However, an important conclusion one can draw from Table 5 results is that both the two features as well as the six features model are pretty stable as compared to their 19-feature counterpart. Even the introduction of news samples from Baby BNC in the training data does not seem to help the predictions of 19 features model. This shows the vulnerability of more complex models to a slight change in the training data.

4 Discussion and Conclusion

In this paper, we have identified two important features that can be very helpful in classifying fiction and non-fiction genres with high accuracy. Fiction articles, *i.e.*, those which are written with an imaginative flavor, tend to have a higher adverb/adjective ratio of POS tags, whereas non-fiction articles, i.e., those which are written in a matter of fact manner, tend to have a higher adjective/pronoun ratio. This not only helps in classification using machine learning but also provides useful linguistic insights. A glance at the percentages of each of these grammatical categories computed over the total number of words in the dataset (Figure 2) reveals several aspects of the genres themselves. In both corpora, the trends are

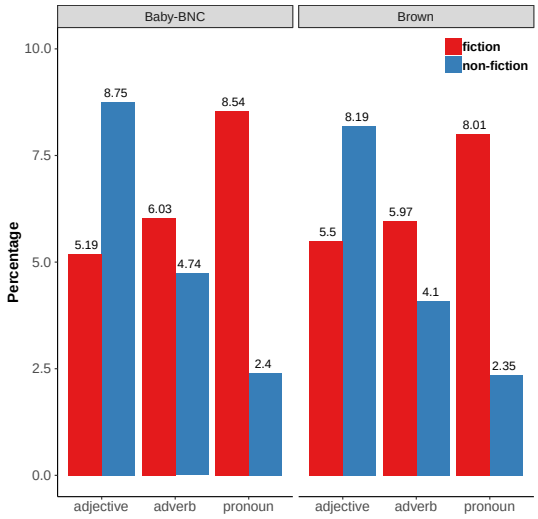

Figure 2: Adjectives, adverbs and pronouns as a percentage of the total number of words

roughly the same. In fiction, both adjectives and adverbs have a roughly similar proportion, while non-fiction displays almost double the number of adjectives compared to adverbs. Also, the percentage of pronouns vary sharply across the two genres in both our datasets as compared to adjectives and adverbs. Figure 3 presents a much more nuanced picture of personal pronouns in the Brown corpus. Fiction displays the greater percentage of third person masculine and feminine pronouns as well as the first person singular pronoun compared to non-fiction, while both genres have comparable percentages of first-person plural *we* and *us*. Moreover, differences in modification strategies using adverbs vs. *wh*-pronouns requires further exploration. Even the usage of punctuation marks differ across genres (Figure 4).

It is worth noting that many guides to writing both fiction (King, 2001) as well as non-fiction (Zinsser, 2006) advise writers to avoid the overuse of both adverbs and adjectives. In a statistical study of classic works of English literature, Blatt (2017) also points to adverb-usage patterns in the works of renowned authors. Nobel prize winning writer Toni Morrison's oft-cited dispreference for adverbs is analyzed quantitatively to show that on an average she used 76 adverbs per 10,000 words (compared to 80 by Hemingway; much higher numbers for the likes of Steinbeck, Rushdie, Salinger, and Wharton). The cited work discusses Morrison's point about eliminating prose like *She says softly* by virtue of the fact that the preceding scene would be described such that

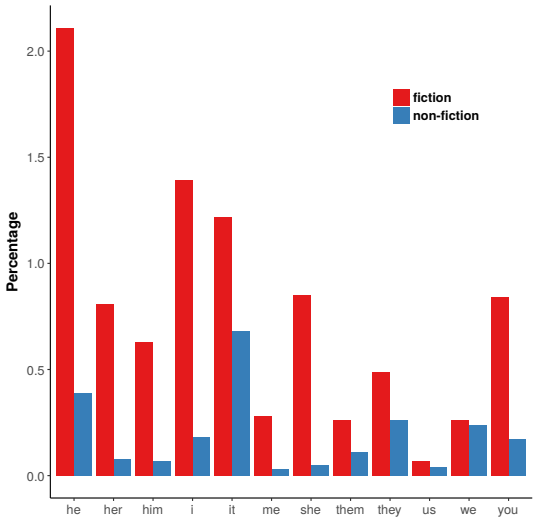

Figure 3: Brown corpus pronouns as a percentage of the total number of words

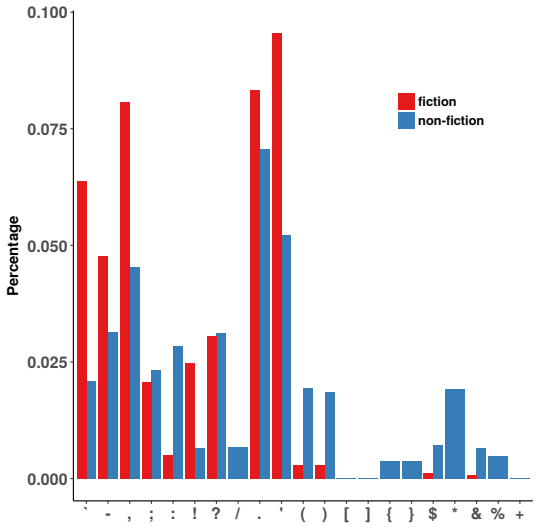

Figure 4: Brown corpus punctuation as a percentage of the number of words

the emotion in the speech is conveyed to the reader without the explicit use of the adverb *softly*. In fact, Sword (2016) advocates the strategy of using expressive verbs encoding the meaning of adverbs as well, as exemplified below (adverbs in bold and paraphrase verbs italicized):

1. She walked **painfully** (*dragged*) toward the car.
2. She walked **happily** (*sauntered*) toward the car.
3. She walked **drunkenly** (*stumbled*) toward the car.
4. She walked **absent-mindedly** (*meandered*) toward the car.

A long line of research undeniably argues that *adjective* and *adverbs* are strong indicators of affective language and serve as an important feature in text classification tasks viz., automatic genre identification (Rittman et al., 2004; Rittman, 2007; Rittman and Wacholder, 2008; Cao and Fang, 2009).In this regard, Rittman and Wacholder (2008) propound that both these grammatical classes have sentimental connotations and capture human personality along with their expression of judgments. For our classifer, rather than the number of adjectives, it is the relative balance of adjectives and adverbs that determine the identity of a particular genre. A large-scale study needs to validate whether this conclusion can be generalized to the English language as a whole. Thus, prescriptions for both technical as well as creative writing should be based on systematic studies involving large-scale comparisons of fictional texts with other non-fiction genres. In particular, the paranoia about elements of modification like adjectives and adverbs seem unjustified as many other mechanisms of nominal and verbal modification like prepositional phrases and subordinate clauses exist in language.[3]

Since our classification is based on the ratios of these POS tags taken across the whole document, it is difficult to identify a few sentences which can demonstrate the role of our features (adverb/adjective and adjective/pronoun ratio) convincingly. Qualitatively, the importance of adjectives can be comprehended with the help of an excerpt taken from the sample file of Brown corpus (*fileid cp09*; adjectives in bold):

" *Out of the church and into his* **big** *car, it tooling over the road with him driving and the headlights sweeping the pike ahead and after he hit college, his expansiveness, the* **quaint** *little pine board tourist courts, cabins really, with a* **cute naked** *light bulb in the ceiling (unfrosted and naked as a streetlight, like the one on the corner where you used to play when you were a kid, where you watched the bats swooping in after the bugs, watching in between your bouts at hopscotch), a room* **complete** *with moths pinging the light and the few* **casual** *cockroaches cruising the walls, an insect Highway Patrol with feelers waving.*"

[3]We are indebted to Mark Liberman's blog post for this idea: https://tinyurl.com/y59jbr64

After removing adjectives (identified using Brown corpus tags), we get:

" *Out of the church and into his car, it tooling over the road with him driving and the headlights sweeping the pike ahead and after he hit college, his expansiveness, the little pine board tourist courts, cabins really, with a light bulb in the ceiling (unfrosted and naked as a streetlight, like the one on the corner where you used to play when you were a kid, where you watched the bats swooping in after the bugs, watching in between your bouts at hopscotch), a room with moths pinging the light and the few cockroaches cruising the walls, an insect Highway Patrol with feelers waving.*"

Although the text with adjectives removed still belongs to the fiction genre, we can clearly see the role that these words can play in enhancing the imaginative quotient of the text. However, counter-intuitively, Figure 2 shows that texts in the non-fiction genre tend to have a higher percentage of adjectives as compared to texts in the fiction genre, but the latter have a higher percentage of adverbs. Hence, this example reiterates the point that the role played by our salient features (adverb/adjective and adjective/pronoun ratios) in classifying fiction and non-fiction genres is difficult to appreciate with only a few lines of text. An interesting question could be to find out the minimum length of a text required for accurate classification into fiction and non-fiction genres and also more significant features in this regard, which we will take up in the future. We also intend to carry out this study on a much larger dataset in the future in order to verify the efficacy of our features.

References

Douglas Biber. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Benjamin Blatt. 2017. *Nabokov's Favourite Word is Mauve: The Literary Quirks and Oddities of Our Most-loved Authors*. Simon & Schuster.

BNC Baby. 2005. *British National Corpus, Baby edition*. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

John F Burrows. 1992. Not unles you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7(2):91–109.

Jing Cao and Alex C Fang. 2009. Investigating variations in adjective use across different text categories. *Advances in Computational Linguistics, Journal of Research In Computing Science Vol*, 41:207–216.

Douglas Douglas. 1992. The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, 26(5-6):331–345.

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.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th conference on Computational linguistics-Volume 2*, pages 1071–1075. Association for Computational Linguistics.

Brett Kessler, Geoffrey Numberg, and Hinrich Schtze. 1997. Automatic detection of text genre. *Proceedings of the 35th annual meeting on Association for Computational Linguistics -*.

S. King. 2001. *On Writing: A Memoir of the Craft*. Hodder & Stoughton.

David YW Lee. 2001. Genres, registers, text types, domain, and styles: Clarifying the concepts and navigating a path through the bnc jungle. *Language Learning & Technology*, 5(3):37–72.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter McCullagh and John A. Nelder. 1989. *Generalized linear models*, volume 37. CRC press, London, New York.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

R.J. Rittman. 2007. *Automatic Discrimination of Genres: The Role of Adjectives and Adverbs as Suggested by Linguistics and Psychology*. Rutgers The State University of New Jersey - New Brunswick.

Robert Rittman and Nina Wacholder. 2008. Adjectives and adverbs as indicators of affective language for automatic genre detection. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, page 65.

Robert Rittman, Nina Wacholder, Paul Kantor, Kwong Bor Ng, Tomek Strzalkowski, and Ying Sun. 2004. Adjectives as indicators of subjectivity in documents. *Proceedings of the American Society for Information Science and Technology*, 41(1):349–359.

Herbert S Sichel. 1975. On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351a):542–547.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 1999. Automatic authorship attribution. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pages 158–164, Stroudsburg, PA, USA. Association for Computational Linguistics.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000. Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 808–814. Association for Computational Linguistics.

Helen Sword. 2016. *The Writer's Diet: A Guide to Fit Prose*. Chicago Guides to Writing, Editing, and Publishing. University of Chicago Press.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1*, pages 173–180. Association for computational Linguistics.

Fiona J Tweedie and R Harald Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352.

W. Zinsser. 2006. *On Writing Well: The Classic Guide to Writing Nonfiction*. HarperCollins.

George Kingsley Zipf. 1932. Selected studies of the principle of relative frequency in language. *Cambridge, Massachusetts: Harvard University Press*.

George Kingsley Zipf. 1945. The meaning-frequency relationship of words. *The Journal of General Psychology*, 33(2):251–256.