Finding a Character's Voice: Stylome Classification on Literary Characters

Liviu P. Dinu, Ana Sabina Uban
Faculty of Mathematics and Computer Science,
Human Language Technologies Research Center,
University of Bucharest
liviu.p.dinu@gmail.com, ana.uban@gmail.com

Abstract

We investigate in this paper the problem of classifying the stylome of characters in a literary work. Previous research in the field of authorship attribution has shown that the writing style of an author can be characterized and distinguished from that of other authors automatically. In this paper we take a look at the less approached problem of how the styles of different characters can be distinguished, trying to verify if an author managed to create believable characters with individual styles. We present the results of some initial experiments developed on the novel "Liaisons Dangereuses", showing that a simple bag of words model can be used to classify the characters.

Keywords: authorship attribution, stylome classification, literary characters, bag of words

1 Previous Work and Motivation

Automated authorship attribution has a long history (starting from the early 20th century (Mendenhall, 1901)) and has since then been extensively studied and elaborated upon. The problem of authorship identification is based on the assumption that there are stylistic features that can help distinguish the real author of a text from any other theoretical author. One of the oldest studies to propose an approach to this problem is on the issue of the *Federalist Papers*, in which Mosteller and Wallace (Mosteller and Wallace, 1963) try to determine the real author of a few of these papers which have disputed paternity. This work remains iconic in the field, both for introducing a standard dataset and for proposing an effective method for distinguishing between the author's styles, that is still relevant to this day, based on function words frequencies. Many other types of features have been proposed and successfully used in subsequent studies to determine the author of a text. These types of features generally contrast with the content words commonly used in text categorization by topic, and are said to be used unconsciously and harder to control by the author. Such features are, for example, grammatical structures (Baayen et al., 1996), part-of-speech n-grams (Koppel and Schler, 2003), lexical richness (Tweedie and Baayen, 1998), or even the more general feature of character n-grams (Kešelj et al., 2003; Dinu et al., 2008). Having applications that go beyond finding the real authors of controversial texts, ranging from plagiarism detection to forensics to security, stylometry has widened its scope into other related subtopics such as author verification (verifying whether a text was written by a certain author) (Koppel and Schler, 2004) or author profiling (extracting information about an author's age, gender, etc).

A related problem that has barely been approached in the scientific literature is that of distinguishing between the writing styles of *fictional* people, namely literary characters. This problem may be interesting to study from the point of view of analyzing whether an author managed to create characters that are believable as separate people with individual styles, especially since style is a feature of speech that is said to be hard to consciously control.

One of the first studies that analyzes literary characters stylistically appeared in John Burrow's "Computation into Criticism" (Burrows, 1987), where he shows that Jane Austen's characters in particular show strong individual styles, which the author distinguishes by comparing lists of the most frequent 30 function words. One more recent study by the same author (Burrows and Craig,

Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Proceedings, pages 78–82, Vancouver, BC, August 4, 2017.

2012) looks at a corpus of seventeenth-century plays and tries to cluster them by character and by playwright, finding in the end that the author signal is stronger than the character one. Another recent study (van Dalen-Oskam, 2014) analyzes the works of two epistolary novels authors, who are known to have written their books together, and tries to solve at the same time the problem of distinguishing between passages written by each author, and between styles of each character in the novel. Using a simple word frequency method to distinguish between the characters, the author finds some of the characters were easier to distinguish than others and concludes that further research is needed.

In this paper we attempt to further the answer to the questions of the best way to solve this problem, and propose some new questions to be approached by future research.

2 Data and Methodology

2.1 Liaisons Dangereuses

The corpus used for this experiment was the 18th century epistolary novel "Liaisons Dangereuses" by Pierre Choderlos de Laclos. The plot of the book is built around two main characters, the Vicomte de Valmont, and the Marquise de Merteuil, who engage with various other characters especially as part of games of seduction, deceit or revenge. The other characters act as their victims, in various roles: Cécile, the innocent young girl who Merteuil plans to corrupt, Danceny, her young passionate admirer, Madame de Tourvel, a faithfull wife who Valmont intends to seduce.

The choice of this text was mainly motivated by the structure of the novel, which is fitting to our problem - as an epistolary novel, it is organized into letters, each written by a different character, which is ideal for easily labelling our text samples with the characters that the text is attributed to. We used the original French version of the text so that we can work on its purest form, unaltered by any noise introduced by translation.

The book consists of 175 letters, sent between the characters; the lengths of the letters vary from 100 to 3600 words, with an average of ~800 words. The routes of the letters sent by and to the main characters can be seen in Table 1 below: the rows correspond to letter senders and the columns to recipients. Table 2 lists the legend for the abbreviations used for the characters' names.

	CV	MM	VV	MV	CD	PT	MR	O
CV		3	2		8			11
MM	1		21	2	2			
VV	2	34			2	12		2
MV		1			1	2	8	
CD	9	3	4	1			2	
PT			9	5			9	
MR				1	1	6		1

Table 1: Letter authors and recipients

Abv.	Character full name
CV	Cécile Volanges
MM	Marquise de Merteuil
VV	Vicomte de Valmont
MV	Madame de Volanges
CD	Chevalier Danceny
PT	Présidente Tourvel
MR	Madame de Rosemonde
O	other

Table 2: Character name legend

2.2 Methodology

We constructed our set of labelled text samples by first splitting the novel into individual letters labelled with their respective "authors". We then only selected the characters who were authors of at least two letters and excluded the rest. We were left with seven main characters: the Marquise de Merteuil, the Vicomte de Valmont, the Présidente de Tourvel, Cécile Volanges, Madame Volanges, the Chevalier Danceny and Madame de Rosemonde. Preprocessing the text involved also removing the first row of each letter, which often contained explicitly the writer and recipient of the letter, so as not to bias the classifier.

3 Text Classification

In order to classify the letters and distinguish between the characters, we used a simple linear support vector machine classifier. We represented the text of the letters starting from a basic bag-of-words model, at first using all words as features in our classifier, then experimenting with additional feature selection, focusing on features that proved to be successful for authorship attribution. In one experiment, we used only content words, using their tf-idf scores as features, after which we tried limiting the features to the k-best features, by using chi^2 feature selection. In another experiment

we tried including only stopwords in the feature set - which are widely used in authorship attribution. Verifying whether these features are still relevant for classifying characters is interesting especially considering they should be harder to consciously manipulate by the author. In a separate experiment, we also tried a feature set of character n-grams, which were previously shown to work for authorship attribution (Dinu et al., 2008), and that are also a very general (and language independent) and versatile type of features that are successfully used in various other natural language processing tasks.

Classification accuracy was measured for each character separately, in a series of leave-one-out experiments. For each character, we built a dataset contaning all letters, from which we excluded at a time one letter written by the target character, to be labelled by our classifier. The dataset was then artificially balanced to contain an equal number of letters pertaining to each character, by only keeping for each character a number of letters equal to the smallest total number of letters written by any character (among the ones we considered). The classification accuracy per character was calculated in the end as the percent of letters written by the character that were correctly classified; the overall accuracy was obtained by averaging the per character accuracy scores (since for character we considered the same number of letters, a simple average results in the overall accuracy).

4 Results and Analysis

Table 3 below illustrates the results (average overall accuracy) for each of the feature sets used, showing that the simple bag-of-words model, including all content words in the text as features, works well for this problem, and additional feature selection do not improve upon these results. The accuracy per character (using the most successful of the models) is shown in Table 4.

This result may look encouraging, as such a simple model is able to obtain a reasonable classification accuracy. On the other hand, it is interesting and worth further investigating that the features demonstrated to work best for authorship attribution do not perform as well on character classification.

We take a closer look at how the characters were classified by showing the confusion matrix containing the misclassified characters, as seen in Table 6. For the same purpose we show an illustration of how the letters, color-coded by author, are grouped together in 2D space, by drawing a scatterplot of the representation of each letter (with content words' tf-idfs as features) after applying 2-dimensional PCA on the feature vectors, shown in Figure 1 below.

Features	Overall accuracy
content words	72.1%
k-best (1000)	69.9%
stopwords	46.6%
char 3-grams	48.5%
char 5-grams	53.3%

Table 4: Overall accuracy for each featureset

Character	Accuracy
CV	95.8%
MM	84.6%
VV	50.0%
MV	75.0%
CD	68.4%
PT	91.3%
MR	55%

Table 5: Classification accuracy per character

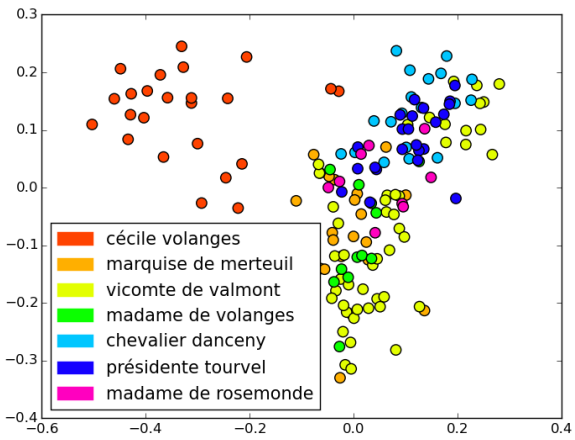

Figure 1: The letters in 2D space of word vector space

Finally, in order to make sense of the importance of each feature for the problem of character classification on our test case, we look at the discriminant features, by taking the list of the highest weighted features from the trained classifier (SVM), shown in table 5 below.

The scatterplot, as well as the confusion matrix, show some interesting insights into how the

Character	Features
CV	aime clef voudrais triste harpe merteuil monsieur petite vicomte maman
MM	sais voudrais merteuil valmont belle harpe aime monsieur danceny maman
VV	présent voudrais aime sais harpe ami amie danceny fille maman
MV	chagrin chose triste clef voudrais harpe amour danceny cécile maman
CD	mal voudrais triste chagrin ami harpe clef aime danceny maman
PT	triste mal aime voudrais harpe ami danceny belle maman neveu
MR	vis faute présidente gercourt danceny madame petite bonne belle vicomte

Table 3: Most discriminating features (bag-of-words)

classifier distinguishes between the letters and the mistakes it makes. In the plot, as well as in the confusion matrix, we can see that the Vicomte de Valmont, the central character of the book, as well as the one involved with most of the other characters, is the character that is hardest to classify. Additionally, he most often gets confused with the Marquise de Merteuil, who is his main interlocutor and "partner in crime". This may point to a common style, but possibly also to common topics of conversation. This hypothesis is enforced by the poor classification results obtained using stopwords as features, as compared to using content words.

	CV	MM	VV	MV	CD	PT	MR
CV	23					1	
MM		22	1			2	1
VV	1	10	26		7	8	
MV				9		1	2
CD	1		1		13	4	
PT			2			21	
MR			1			3	5

Table 6: Confusion matrix for character classification

5 Conclusions and Future Directions

We have shown that a simple bag of words model using a linear support vector machine as a classifier can be used to distinguish between characters of a literary work. It is unclear though whether the classifier captures style in the same sense as in authorship attribution, or rather characters' prefferred ideas or topics of conversation for example. From this point of view it may be interesting to compare these results to a topic modelling approach on the same dataset, as well as further explore the attributes of the most useful features.

In the future it may also be interesting to look at how various authors pertaining to different periods and literary currents compare in terms of their ability (and desire) to create individual, stylistically independent characters. Literary theory (Wellek and Warren, 1956) tells us that the practice of giving characters strongly individual voices is a rather modern idea, which may be interesting to confirm experimentally. In theory, literary characters evolved with time and literary current from the classical figures, who represented a typology, to the realist characters, who are pictured with strong individualities.

Further, the analogous problem to author profiling could be tackled with regard to literary characters. Separately of whether characters are easy to distinguish stylistically from one another, it may be interesting to see if an author managed to belivably build a character's style that is consistent with features of the character's personality: such as age or gender. Can older authors write from the point of view of teenagers (a notable example of this is Salinger's *Catcher in the Rye*), can males create consistent female characters? These are questions that we intend to approach in further experiments on this topic.

References

Harald Baayen, Hans Van Halteren, and Fiona Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11(3):121–132.

John Burrows and Hugh Craig. 2012. Authors and characters. *English studies* 93(3):292–309.

John Frederick Burrows. 1987. *Computation into criticism: A study of Jane Austen's novels and an experiment in method*. Clarendon Pr.

Liviu Petrisor Dinu, Marius Popescu, and Anca Dinu. 2008. Authorship identification of romanian texts with controversial paternity. In *LREC*.

Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*. volume 3, pages 255–264.

Moshe Koppel and Jonathan Schler. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*. volume 69, page 72.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*. ACM, page 62.

Thomas Corwin Mendenhall. 1901. A mechanical solution of a literary problem. *Popular Science Monthly* .

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association* 58(302):275–309.

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.

Karina van Dalen-Oskam. 2014. Epistolary voices. the case of elisabeth wolff and agatha deken. *Literary and Linguistic Computing* page fqu023.

Rene Wellek and Austin Warren. 1956. *Theory of literatures*. Harcourt, Brace & World New York.