Stopwords and Stylometry: A Latent Dirichlet Allocation Approach

Arun R.
arun_r@csa.iisc.ernet.in

Saradha R.
saradha.ravi@gmail.com

V. Suresh
vsuresh@csa.iisc.ernet.in

M. Narasimha Murty
mnm@csa.iisc.ernet.in

C. E. Veni Madhavan
cevm@csa.iisc.ernet.in

Department of Computer Science and Automation
Indian Institute of Science
Bangalore 560 012, INDIA

Abstract

We illustrate the utility of generative models for the purpose of stylometry – the science of author attribution. Though content words provide semantic handles and intuitively relate to author-styles, they are usually associated with a large vocabulary and are not consistent across corpora. On the contrary, stopwords are limited in number and do not suffer from the above mentioned issues and yet seem to retain abstract signatures of author style. We explore the use of Latent Dirichlet Allocation on stopwords and show that the resulting topic distributions provide robust handles to classify authors and help perform authorship attributions. In addition to this, we also observe that they are effective in identifying the gender of the authors.

1 Introduction

Stylometry is the collective name given for the approaches that are used to study author styles. A quantitative analysis of authors' styles would help infer the authorship of anonymous documents and as a digital copyrights aide. In 1962, Mosteller and Wallace[1] presented the first machine learning approach for stylometry with which they resolved the authorship of the disputed articles of the Federalist Papers (http://en.wikipedia.org/wiki/Federalist_Papers). Since then stylometry has benefitted by the use of feature based machine learning and pattern classification methods[2]. Interestingly, Mosteller and Wallace[1] observed that it is the distribution of stopwords in the text rather than the meaningful content words that provided handles for their classification – This observation seems to be enduring in that we showed in our recent work[3] the effectiveness of stopwords and their interactions defined in terms of the gaps between them for multi-author classification. In this work, we show the effectiveness of Latent Dirichlet Allocation (LDA)[4] in performing author classification and in identifying the authors' gender.

2 Contribution of the Paper

Typically LDA is used for for extracting common topics from text corpora with the view to find hidden or latent topic structures in them in the form of probability distributions of words over topics and topics over documents. These distribution vectors are subsequently used for classification.

To our knowledge, this is the first time use of LDA for the purpose of stylometry. In this regard, we present to LDA streams of stopwords[5] that result from stripping away of the content words from normal text. Intuitively, the resulting topics do not have any semantic meaning as is usually observed with topic discerned from content words. Our primary contribution is the observation that the abstract topics and their distributions are nevertheless meaningful quantifiers for stylometry. In addition to identifying authors, we also show that our approach is effective in identifying the gender of the author.

This paper is organised as follows. The next section describes the data, the preprocessing done on them and experiments we have performed. We then describe our results on multi-author classification and gender identification in section 4. Our concluding remarks are presented in section 5.

3 Data

The data used for our experiments are sourced from the gutenberg repository at www.gutenberg.org. The literary works of 12 different authors were used for conducting experiments for multi-author classification problem. For these experiments, the data was used to provide three different kinds of inputs: a. Text constituted by the stop words only, b. The text constituted by content words, and c. The entire text of the novel – stopwords plus the content words. The number of unique content words are in the range $\sim 100,000$ whereas the number of stopwords considered number only 555. The authors and statistics of the text used for our experiments are given in Table 1. As can be seen from it, the data is a careful mix of literary genres and are from both genders so as to avoid any biases in the classification. In addition to these precautions, the time-span of the texts considered are spread over a period of 100 years in order to avoid biases that might inadvertently result from the idiosyncrasies of narrow time-spans.

Author	Genre[1]	Timeline	Stopwords	Content Words
Daniel Defoe	A,F,Hi	1808-1894	477201	210550
Jane Austen	F,Hu,Ps	1811-1818	474305	248676
Allen Grant	B,F,Sc,Ph	1848-1899	215001	148670
George Elliot	F,Ps	1859-1871	520661	311030
Harold Bindloss	Ro,F	1866-1945	525724	321902
James Otis	A,C,F,Hi	1883-1899	252518	136089
George Bernard Shaw	D,F,Hi,Hu,W	1885-1912	145385	85476
Hamlin Garland	A,F,Ps,Ro,Sp,T	1897-1921	349296	229164
Captain Ralph Bonehill	A	1902	175659	105169
Phillips Oppenheim	F,M,Po	1902-1920	415892	243386
G K Chesterton	M,Ph,Ps,Re	1905-1916	186409	111290
Baronness Orczy	A,Hi,Po,Ro	1905-1921	392127	261019
Total	18	1808-1921	4130178	2307252

Table 1: Statistics of the data used for the experiments

[1]List of Genres considered: A - Adventure, Au - Autobiography, B - Biography, C - Children, D - Drama, F - Fiction Hi - History, Hu - Humour, M - Mystery, Ph - Philosophy, Po - Politics, Ps - Psychology R - Religion, Ro - Romance, Sc - Science, Sp - Spirituality, T - Travel, W - War

4 Experiments and Results

The experiment involved the following steps. First, the works of the authors is split into documents of 5000 words – This number was varied and found that 5000 is a reasonable size for the training and testing documents from the classifier's accuracy point of view. Also this number is at most 1% of the number of words per author and hence illustrative of the classifier's potential to generalize based on smaller input sizes. Next, LDA is used to generate the topic-probabilities for every document, for all the authors. We varied the number of topics for LDA over the range $5 - 50$ and found that 25 topics correspond to the best classification results. The LDA method used Gibbs sampling and had the values of the hyper-parameters α and β to be 1.0 and 0.1 respectively. The LDA implementation that was used in our experiments is available at http://gibbslda.sourceforge.net/

Once the topic vectors for each document is available we use libSVM – an SVM implementation available at http://www.csie.ntu.edu.tw/ cjlin/libsvm – for the author classification task. LibSVM's default values were retained as the SVM parameters in our experiments. For training, the document count was varied on three counts: 10, 20 and 30, for all the authors. For each count, the remaining documents were used for testing. Each author in our database had on the average 50 documents. Hence at least 20 documents were available for testing purposes. The experiments were performed for the three different kinds of inputs as mentioned earlier — stopwords only, content words only and stopwords + content words. The results are shown in table 2. Best results are observed with 25 topics are higher for stopwords only category when compared to that of the other two. It is interesting to note that in comparison to the categorises that take into account content words, similar or higher accuracy in classification is obtained with stopwords alone. This needs to viewed in terms of the disparity in the vocabulary sizes involved — stopwords: 555 and content words: $100,000$. This is illustrative of LDA's ability to generalize based on a small vocabulary.

The table also shows for comparison, classification based on Naive Bayes on documents of 5000 stopwords to contrast the classification accuracies obtained with LDA. The individual stopword frequencies serve as features for the Naive Bayes classifier. The results are comparable if the document size is increased to to include a significant portion of the authors' contribution, say, $100,000$ stopwords. This should be seen in the context of similar accuracy obtained for just 5000 stopwords using LDA. Note that Naive Bayes is not practical for content word frequencies as the training document sizes would have to be unacceptably large for computing any meaningful statistics.

4.1 Federalist Papers

We now present our results on the Federalist papers problem which was originally solved by Mosteller and Wallace[1]. We observe that, even with just two topics, all the 12 out of the 12 articles of the disputed Federalist Papers are in line with the generally agreed results. On the other hand, with just two topics for content words only streams we get an accuracy of 5 out of 12. Increasing the number of topics to 10 for content words improves it to 9 out of 12. Content+stopwords with two topics gives an accuracy of 9 out of 12 which improves to 12 out of 12 with 10 topics. In addition to the positive results from the above multi-author classification, this is ample illustration that stopwords in conjunction with LDA contribute effective features for stylometry.

A fact to be noted here is that in some of the Federalist Papers, the total number of stopwords do not add up to even 5000 as we have required in the above 12 author classification. The correlation between the number of authors simultaneously classified and the number of topics considered needs further study.

4.2 Gender Identification

Our database consists of three female authors and nine male authors. We are interested in testing the potential of the topic vectors generated from stopword streams for gender classification. The topic vectors generated for author identification purposes are now grouped into two classes *male* and *female*. We consider 10 to 40 documents from each class for training and the remaining vectors were tested for the accuracy of classifying the authors' gender. The number of topics considered are 25. We observe a classifier accuracy in the range of $75 - 80\%$ for around 750 documents. We note that the classification was made more difficult by not obtaining the training documents through a uniform sampling over all available documents from all authors; instead they were derived from one

(from training vectors numbering 10) or at the maximum two authors for training vectors upto 40. Also the authors of the training documents were not used for testing. Thus the topic vectors from a any single male author is a good enough generalization to distinguish topic vectors resulting from any female author in our database.

Words considered	Training	5 topics	15 topics	25 topics	50 topics	NAIVE BAYES
STOP WORDS	10	54.2	71.71	82.63	77.6	47.91
VOCAB SIZE: 555	20	56.39	89.23	86.03	85.52	30.96
	30	56.12	89.03	92.2	89.24	19.72
CONTENT WORDS	10	54.2	72.99	73.69	59.93	
VOCAB SIZE: 93425	20	56.4	72.46	76.43	66.07	
	30	80.12	93.99	85.75	78.64	
ALL WORDS	10	56.28	82.57	70.02	69.85	
VOCAB SIZE: 93980	20	58.04	81.81	68.98	70.22	
	30	60.17	81.14	67.68	72.55	

Table 2: Percentage classification accuracies for the experimental results

5 Conclusions

We have applied LDA on stopword streams and demonstrated its efficacy in author and author-gender identification over a database of novels that span a wide time-line and a good mix of genres. Our results indicate that stopwords in conjunction with LDA are robust features for stylometric purposes. Since our approach uses stopwords as features, it places minimal demands on the number of words required for authorship attribution. The abstract topic distributions over stopwords assigned by the LDA mechanism seems to be as meaningful as the intuitive and semantically relatable topic distribution over content words. As seen from the table 1 the data used spans across genres and has a good mix of gender; our approach performs well across this mix. Identifying features that are sensitive towards genre and the time-period of the documents would be among our topics of interests for future studies.

References

[1] Mosteller, F. & Wallace, D. L. (1984) *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, Springer-Verlag.

[2] Juola, P. (2006) Authorship attribution. *Found. Trends Inf. Retr.*, 1(3):233–334.

[3] R. Arun, V. Suresh & C. E. Veni Madhavan (2009) Stopword Graphs and Authorship Attribution in Text Corpora. *Third International Conference on Semantic Computing*, 192-196

[4] Blei, D., Ng, A. & Jordan, M. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–022.

[5] Lewis, D. D., Yang, Y., Rose, T. G., Li, F. (2004) Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.