Authorship Attribution Datasets by Numbers

Each dataset is gathered from different sources such as web blogs, twitter and newspapers. A shallow analysis of the topic, source and author distributions is given in Table 3. Statistics for sampled datasets with specified numbers of authors are obtained for the union of 10 test datasets generated by 10-fold stratified cross-validation. In Table 1, datasets are described by L for large, S for small, and Top-k for the top-k authors. Author count is the number of distinct authors, document count is the number of documents, and token count is the number of distinct tokens in all documents. Average (Avg.) values are computed via Eq. 26, Eq. 27, Eq. 28 and Eq. 29. We obtained topic models by Latent Dirichlet Allocation (LDA) with a topic size of three on the complete datasets and used these models to analyze the training and testing datasets.

$$Avg.authors\ per\ topic = \frac{\sum_{k}^{authors} number\ of\ topics\ written\ by\ author_k}{number\ of\ authors} \quad (26)$$

$$Avg.topics\ per\ author = \frac{\sum_{p}^{topics} number\ of\ authors\ write\ in\ topic_p}{number\ of\ topics} \quad (27)$$

$$Avg.authors\ per\ source = \frac{\sum_{k}^{authors} number\ of\ sources\ written\ by\ author_k}{number\ of\ authors} \quad (28)$$

$$Avg.sources\ per\ author = \frac{\sum_{s}^{sources} number\ of\ authors\ write\ in\ source_s}{number\ of\ sources} \quad (29)$$

Table 1. Dataset statistics

Dataset	Author count	Document count	Token count	Avg. authors per topic	Avg. topics per author	Avg. authors per source	Avg. sources per author
PAN11-L Train	72	10635	32069	35	2.430	-	-
PAN11-L Test	72	1300	9424	16.8	1.412	-	-
PAN11-S Train	26	3519	14936	9.6	1.846	-	-
PAN11-S Test	26	495	4778	12.5	1.186	-	-
PAN18-1 Train	20	140	10480	-	-	-	-
PAN18-1 Test	20	105	8235	-	-	-	-
PAN18-2 Train	5	35	4662	-	-	-	-
PAN18-2 Test	5	21	3255	-	-	-	-
PAN18 1&2 Train	25	175	10480	-	-	-	-
PAN18 1&2 Test	25	126	8235	-	-	-	-
C10 Train	10	500	15467	-	-	-	-
C10 Test	10	500	15737	-	-	-	-
C50 Train	50	2500	36845	50.0	3.0	-	-
C50 Test	50	2500	37914	45.6	2.7	-	-
PAN11 Top-5	5	600	5551	5.0	3.0	-	-
PAN11 Top-10	10	1200	8767	10.0	3.0	-	-
Articles Top-5	5	600	25242	5.0	3.0	-	-
Articles Top-10	10	1200	46228	8.6	2.6	-	-
Blogs Top-5	5	600	48718	5.0	3.0	-	-
Blogs Top-10	10	1200	94764	7.1	2.2	-	-
Tweets Top-5	5	600	3794	5.0	3.0	-	-
Tweets Top-10	10	1200	6255	9.6	2.9	-	-
Articles & Blogs	10	1200	50666	8.6	2.6	5.0	1.0
Articles & Tweets	10	1200	27070	8.6	2.6	8.5	1.7
Blogs & Tweets	10	1200	30723	9.66	2.9	7.5	1.4
Articles & Blogs & Tweets	10	1200	36191	10.0	3.0	6.33	1.9

In Table 1, the token count represents the number of unique tokens in the dataset. Token counts depend on the topic diversity, the number of documents and the token lengths of the documents in the dataset. All datasets have balanced topic distributions and most of the documents are written on a single topic. The distributions of topics among authors and author topic tendencies are effective measures in the analysis of performance. Table 1 provides a shallow insight on these topic distributions; for example, in the C50 dataset, all authors are associated with documents on all topics.

In Figure 1, a shallow analysis of the topic distributions for each author is presented for sampled datasets. A, T, and S denote author, topic and source, respectively. In addition to topic distributions, source distributions for each author are included for multisource datasets (except Article & Blogs because in the Article & Blogs dataset, the authors write either articles or blogs). Source abbreviations are given in the datasets by order of appearance. For example, in the Article & Tweets datasets, S0 corresponds to articles and S1 corresponds to tweets. The main objective of these visualizations is to facilitate understanding of the topical clusters and the distribution of sources to determine whether a correlation exists with the AA scores. To the best of our knowledge of the sources and authors, these topics fall into the subcategories of politics, sports and technology. More accurate topic models can be obtained by using external sources such as Wikipedia; however, this is outside the scope of this analysis.

Figure 1. Topic and source distributions for authors in sampled datasets