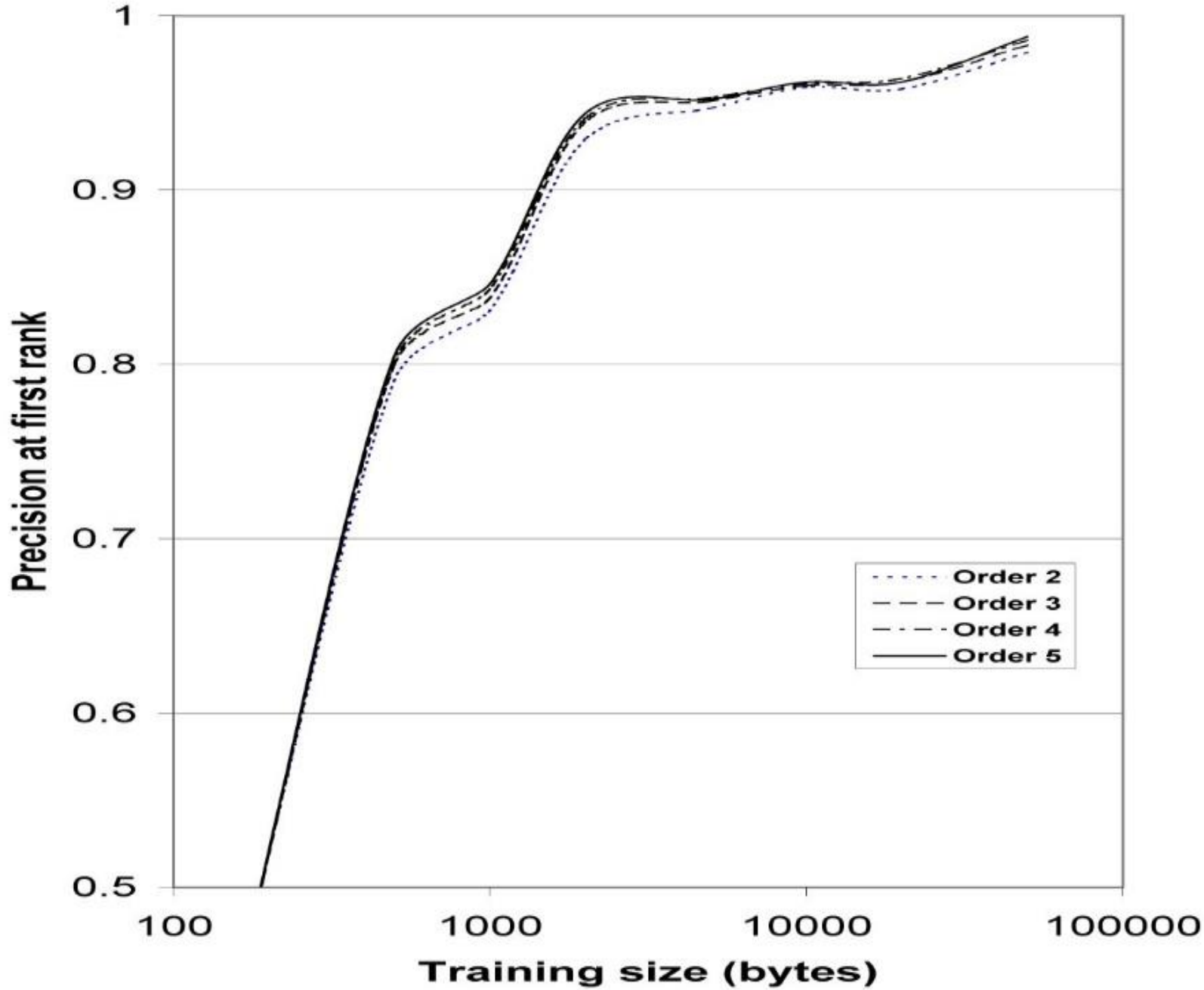
William J. Teahan

Using compression-based language models for text categorization

Jakob Köhler

$$H(p_M, D) = -\frac{1}{n} \log_2 p_M(D), \quad D = x_{1n}$$


1
0.9
0.8
0.7
0.6
0.5
Precision at first rank
100
1000
10000
100000
Training size (bytes)
Order 2
Order 3
Order 4
Order 5

$$
\begin{aligned}
&H(p_M, D) \\
&\quad = -\tfrac{1}{n} \log_2 p_M(D), \quad D = x_{1n} \\
&\quad = -\tfrac{1}{n} \log_2 \textstyle\prod_{i=1}^{n} p_M(x_i | context_i) \quad [by\ Chain\ Rule] \\
&\quad = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} -\log_2 p_M(x_i | context_i)
\end{aligned}
$$

Disputed papers

No.	Madison (bpc)	Hamilton (bpc)	No.	Madison (bpc)	Hamilton (bpc)
49	**1.79**	1.93	55	**1.86**	1.97
50	**1.92**	2.07	56	**1.70**	1.85
51	**1.72**	1.88	57	**1.85**	1.98
52	**1.78**	1.94	58	**1.80**	1.93
53	**1.79**	1.93	62	**1.83**	1.84
54	**1.73**	1.89	63	**1.82**	**1.82**

Papers known to have been written by Madison

Papers known to have been written by Hamilton

No.	Madison (bpc)	Hamilton (bpc)	No.	Madison (bpc)	Hamilton (bpc)
44	**1.76**	1.90	59	**1.82**	1.88
45	**1.67**	1.81	60	1.78	**1.71**
46	**1.78**	1.85	61	1.78	**1.73**
47	**1.70**	1.81	65	1.87	**1.82**
48	**1.85**	1.99	66	1.80	**1.75**

Table 1.3. Ascribing authorship to the Federalist Papers.