Table 5. F1 scores for mention detection and KORE50 grouped by annotator and domain, portraying domain-specific performance for each system. Dash ("-") means no available entry, e.g. for datasets when no topic of that domain was detected.

	KORE 50											
System	MED	POL (2)	FINMA	GOV (2)	ANALYSIS (2)	PLAYERS	CORP (1)	CELEBNEWS (9)	CHAMP	EVENT (3)	MATCHES (30)	GRMN (1)
Babelfy	-	0.5077	-	0.5179	0.3429	-	0.6154	0.4007	-	0.3491	0.4314	0.8000
CLOCQ	-	0.8333	-	0.5833	0.8333	-	0.5714	0.6019	-	0.7556	0.6835	0.6667
DBp. Spotlight	-	0.7222	-	0.4762	0.6500	-	0.6000	0.6196	-	0.5037	0.5321	0.8000
Falcon 2.0	-	0.0000	-	0.5333	0.0000	-	0.0000	0.2984	-	0.1905	0.1532	**1.0000**
FLAIR	-	**1.0000**	-	0.8333	**1.0000**	-	0.7500	**0.8704**	-	**1.0000**	**0.9560**	**1.0000**
OpenTapioca	-	0.5000	-	0.0000	0.2500	-	0.0000	0.1000	-	0.2667	0.0935	0.6667
REL	-	**1.0000**	-	**1.0000**	**1.0000**	-	0.7500	0.8426	-	**1.0000**	0.9451	**1.0000**
ReFinED	-	**1.0000**	-	0.8333	**1.0000**	-	**1.0000**	0.8481	-	0.9167	0.8886	**1.0000**
ReLiK	-	0.8333	-	0.4000	**1.0000**	-	0.8571	0.6397	-	0.7556	0.7432	0.6667
spaCy	-	0.9286	-	0.5333	0.1429	-	0.8571	0.6296	-	0.9524	0.7345	0.6667
TagMe	-	0.7333	-	0.5397	0.7750	-	0.8000	0.6273	-	0.4272	0.6757	0.8000
TextRazor	-	**1.0000**	-	0.7500	0.5000	-	0.6667	0.5735	-	0.5833	0.8118	**1.0000**

A. Appendices

A.1. Fine-Grained Dataset Tables

We add tables with fine-granular domain-grouped results for each dataset to the appendix, so that researchers interested in in-detail findings outside the possible scope for this paper's main body may take a look at them. We provide our tables displaying domain-specific results for each singular dataset and identified reduced set of domains. Further, the number of documents a displayed primary topic represents follows the abbreviation between brackets to see how influential a given domain is for the specific dataset. KORE50 (Table 5) covers all domains other than MED, FINMA, PLAYERS and CHAMP, with most documents being included within MATCHES or CELEBNEWS.

Reuters-128 (Table 6) is mostly made up identified FINMA and EVENT documents with 5 or fewers documents classified as MED, POL, GOV and CHAMP.

As for News-100 (Table 7), it is a dataset consisting of German news documents and our topic model correctly identifies all of its documents as being part of the single GRMN category.

The RSS-500 (Table 8) dataset comprised of newspaper RSS feeds spans all identified topics other than PLAYERS with most documents being recognised as part of GOV, CORP and CHAMP with EVENT, POL and ANALYSIS tying with 28 documents each.

AIDA-CoNLL (Table 9) is the biggest general-domain dataset we include in this paper with most documents being identified as part of POL, FINMA, PLAYERS, ANALYSIS and CELEBNEWS.

Finally, MedMentions (Table 10), a dataset comprised of medical and clinical data is correctly identified as being majorly within the MED domain with 4341 documents.

A.2. Statistical Significance

To support our claim about the existence of domain bias in annotators, we performed a p-value analysis comparing F1 scores computed by systems using a uniform sampling strategy, and F1 scores by annotators using trained machine learning models presented in this paper. The p-value measures the probability of observing (at least) the current results, if the null hypothesis were true. In the setup of this paper, the null hypothesis H_0 claims that there is no domain bias in annotation systems. Therefore, using a uniform recommender would yield the same F1 score as a topic-based recommender. The alter-

Table 6. F1 scores for REUTERS grouped by annotator and domain, portraying domain-specific performance for each system. Dash ("-") means no available entry, e.g. for datasets when no topic of that domain was detected.

	REUTERS											
System	MED (2)	POL (5)	FINMA (81)	GOV (1)	ANALYSIS	PLAYERS	CORP	CELEBNEWS	CHAMP (2)	EVENT (36)	MATCHES	GRMN
Babelfy	0.0298	0.1088	0.0740	0.1333	-	-	-	-	0.0000	0.0805	-	-
CLOCQ	0.0000	0.0000	0.3169	0.1333	-	-	-	-	0.3333	0.2912	-	-
DBp. Spotlight	0.0871	0.1585	0.0954	0.2000	-	-	-	-	0.0588	0.1390	-	-
Falcon 2.0	0.0000	0.0000	0.0000	0.0000	-	-	-	-	0.0000	0.0000	-	-
FLAIR	**0.4313**	0.6991	**0.8072**	**1.0000**	-	-	-	-	**1.0000**	**0.7701**	-	-
OpenTapioca	0.0667	0.4564	0.1254	0.0000	-	-	-	-	0.0000	0.1225	-	-
REL	0.4206	**0.7011**	0.5445	0.6667	-	-	-	-	0.1667	0.4217	-	-
ReFinED	0.3742	0.4920	0.4380	0.6667	-	-	-	-	0.5000	0.6389	-	-
ReLiK	0.3718	0.6931	0.4257	0.6667	-	-	-	-	0.0000	0.3055	-	-
spaCy	0.2524	0.4435	0.3672	0.6667	-	-	-	-	0.5000	0.5221	-	-
TagMe	0.0765	0.2126	0.1058	0.2000	-	-	-	-	0.0435	0.1233	-	-
TextRazor	0.1432	0.4155	0.2641	0.6667	-	-	-	-	0.5096	0.4396	-	-

Table 7. F1 scores for mention detection and NEWS grouped by annotator and domain, portraying domain-specific performance for each system. Dash ("-") means no available entry, e.g. for datasets when no topic of that domain was detected.

	NEWS											
System	MED	POL	FINMA	GOV	ANALYSIS	PLAYERS	CORP	CELEBNEWS	CHAMP	EVENT	MATCHES	GRMN (100)
Babelfy	-	-	-	-	-	-	-	-	-	-	-	0.1771
CLOCQ	-	-	-	-	-	-	-	-	-	-	-	0.1111
DBp. Spotlight	-	-	-	-	-	-	-	-	-	-	-	0.0313
Falcon 2.0	-	-	-	-	-	-	-	-	-	-	-	0.0022
FLAIR	-	-	-	-	-	-	-	-	-	-	-	0.2178
OpenTapioca	-	-	-	-	-	-	-	-	-	-	-	0.1693
REL	-	-	-	-	-	-	-	-	-	-	-	0.2845
ReFinED	-	-	-	-	-	-	-	-	-	-	-	0.4581
ReLiK	-	-	-	-	-	-	-	-	-	-	-	**0.6217**
spaCy	-	-	-	-	-	-	-	-	-	-	-	0.2162
TagMe	-	-	-	-	-	-	-	-	-	-	-	0.1818
TextRazor	-	-	-	-	-	-	-	-	-	-	-	0.5320

Table 8. F1 scores for mention detection and RSS grouped by annotator and domain, portraying domain-specific performance for each system. Dash ("-") means no available entry, e.g. for datasets when no topic of that domain was detected.

	RSS											
System	MED (1)	POL (28)	FINMA (10)	GOV (170)	ANALYSIS (28)	PLAYERS	CORP (98)	CELEBNEWS (9)	CHAMP (80)	EVENT (28)	MATCHES (28)	GRMN (1)
Babelfy	0.1053	0.1056	0.0893	0.0827	0.1821	-	0.1770	0.1603	0.0670	0.1238	0.1454	0.1429
CLOCQ	0.0000	0.2816	0.3406	0.3003	0.3862	-	0.3983	0.3304	0.2923	0.2634	0.4147	0.5000
DBp. Spotlight	**0.1538**	0.1670	0.1421	0.1866	0.2767	-	0.2740	0.2813	0.1423	0.2007	0.3172	0.1818
Falcon 2.0	0.0000	0.0000	0.0000	0.0012	0.0000	-	0.0000	0.0000	0.0073	0.0000	0.0000	0.0000
FLAIR	0.0000	**0.4884**	**0.7157**	**0.6548**	**0.5793**	-	**0.6828**	0.6360	**0.7102**	**0.6524**	**0.7172**	**1.0000**
OpenTapioca	0.0000	0.1745	0.2067	0.1603	0.2982	-	0.2527	0.3148	0.1530	0.1036	0.1952	0.0000
REL	0.0000	0.3643	0.4267	0.4101	0.5659	-	0.5910	0.6265	0.3860	0.4422	0.5942	**1.0000**
ReFinED	0.0000	0.4474	0.6631	0.5878	0.4578	-	0.5980	0.5571	0.6142	0.6111	0.6167	0.8000
ReLiK	0.0000	0.3895	0.3638	0.3474	0.5697	-	0.5183	**0.6591**	0.2891	0.4286	0.5024	**1.0000**
spaCy	0.0000	0.4097	0.6727	0.5370	0.4488	-	0.5463	0.4182	0.5515	0.5294	0.6234	0.8000
TagMe	0.0000	0.1536	0.1546	0.1730	0.2425	-	0.2861	0.3146	0.1405	0.1974	0.3036	0.4444
TextRazor	0.0000	0.4293	0.4724	0.5348	0.5362	-	0.5545	0.5231	0.5883	0.5003	0.5681	0.6667

Table 9. F1 scores for mention detection and AIDA-CoNLL grouped by annotator and domain, portraying domain-specific performance for each system. Dash ("-") means no available entry, e.g. for datasets when no topic of that domain was detected.

	AIDA-CoNLL											
System	MED (5)	POL (492)	FINMA (275)	GOV (11)	ANALYSIS (204)	PLAYERS (215)	CORP (6)	CELEBNEWS (143)	CHAMP	EVENT (36)	MATCHES (1)	GRMN
Babelfy	0.1023	0.2178	0.1370	0.2204	0.3286	0.4883	0.2531	0.3677	-	0.1512	0.2687	-
DBp. Spotlight	0.1279	0.3054	0.1888	0.3914	0.4513	0.6570	0.4111	0.4891	-	0.2196	0.3077	-
Falcon 2.0	0.0000	0.0020	0.0022	0.0000	0.0147	0.0245	0.0000	0.0167	-	0.0000	0.0000	-
FLAIR	0.6533	0.8550	0.7011	0.9028	0.8989	0.8551	0.9192	0.9231	-	0.6544	0.8462	-
OpenTapioca	0.2403	0.3506	0.2189	0.1953	0.4087	0.3456	0.2580	0.3172	-	0.2276	0.5000	-
REL	0.6524	0.8176	0.6179	0.8159	0.7900	0.7581	0.8192	0.7359	-	0.6576	0.8696	-
ReFinED	0.3450	0.6355	0.3462	0.7098	0.6235	0.6702	0.6144	0.6844	-	0.4837	0.6875	-
ReLiK	**0.9069**	**0.9726**	**0.9495**	**0.9971**	**0.9648**	**0.9660**	**0.9685**	**0.9689**	-	**0.9284**	**1.0000**	-
spaCy	0.2350	0.5500	0.2668	0.6223	0.4547	0.2835	0.4253	0.4441	-	0.4073	0.4828	-
TagMe	0.1279	0.2551	0.1687	0.2557	0.3971	0.6371	0.3230	0.4016	-	0.1851	0.3043	-
TextRazor	0.1960	0.4838	0.2570	0.5517	0.5572	0.4552	0.5509	0.6008	-	0.3465	0.5143	-

Table 10. F1 scores for mention detection and MedMentions grouped by annotator and domain, portraying domain-specific performance for each system. Dash ("-") means no available entry, e.g. for datasets when no topic of that domain was detected.

	MedMentions											
System	MED (4341)	POL (7)	FINMA (2)	GOV (2)	ANALYSIS (1)	PLAYERS	CORP	CELEBNEWS	CHAMP (2)	EVENT (2)	MATCHES	GRMN
Babelfy	0.0483	0.0882	0.0857	0.1969	0.0526	-	-	-	0.1875	0.0821	-	-
CLOCQ	**0.0791**	0.0000	0.0000	0.0000	0.0000	-	-	-	0.0000	0.0000	-	-
DBp. Spotlight	0.0464	0.1006	**0.1076**	0.1750	0.0870	-	-	-	0.1833	0.0465	-	-
FLAIR	0.0054	0.0603	0.0109	0.0714	0.0000	-	-	-	0.0000	0.0000	-	-
OpenTapioca	0.0014	0.0106	0.0000	0.0769	0.0000	-	-	-	0.0000	0.0000	-	-
REL	0.0064	0.0530	0.0108	0.0769	0.0000	-	-	-	0.0000	0.0000	-	-
ReFinED	0.0044	0.0000	0.0099	0.0667	0.1250	-	-	-	0.0435	0.0000	-	-
ReLiK	0.0063	0.0281	0.0110	0.0769	0.0000	-	-	-	0.0000	0.0000	-	-
spaCy	0.0084	0.0499	0.0100	0.0667	0.0000	-	-	-	0.0000	0.0000	-	-
TagMe	0.0613	**0.1523**	0.0000	**0.2241**	**0.1379**	-	-	-	**0.2185**	0.0500	-	-
TextRazor	0.0459	0.0784	0.0189	0.0625	0.0800	-	-	-	0.0940	**0.0868**	-	-

Doc. Embedding

RF	SVM	k-NN	MLP
7.4309e-45	1.6015e-47	2.4737e-44	1.4689e-46

Primary Topic (1-Hot)

RF	SVM	k-NN	MLP
1.0900e-47	6.4099e-51	1.2041e-47	9.9667e-51

Primary Topic & Doc. Embeddings

RF	SVM	k-NN	MLP
3.4078e-39	1.5211e-40	3.2825e-42	3.4844e-40

Table 11. Computed p values for our experiments for result significance.

native hypothesis H_1 states that domain bias exists in annotators, with the topic-based recommender achieving significantly higher F1 scores throughout experiments. The low p-values, shown in Table 11, provide a strong statistical evidence for rejecting the null hypothesis H_0 and confirms an almost certain existence of of domain bias in annotation systems H_1.

A.3. Limitations & Further Notes

Our approach generates a *best-vs-all* type of annotation dataset ranking on a document-level based on F1 scores. We experimented with applying a softened gap threshold of 5% and noticed no noticeable difference for annotation recommendation. Despite our choice of focusing F1, there are valid arguments to account for precision or recall instead, depending on us case. Further, the design of *best-vs-all* comes with some disadvantages regarding recommendation, such as not considering lower-ranking systems outside of defined ranges. Theoretically, these could be consistently barely below the best-ranking annotation system, but still overtake all other ones on average – our experiments with our current datasets however seem to indicate that it was not a significant edge case to the point of greatly affecting performance.

Due to the nature of the problem we are trying to solve, it is likely for there to be duplicate best systems for a given document. As such, we generate multiple labels for the same datapoint to a non-insignificant amount, generalising, but also potentially confusing our model due to the similitude of the input signals expecting varying outputs and forcibly

dragging recommendation results down. Further, despite having put considerable effort into our choice of systems, it would be great to have more specialised domain-specific linking approaches to make use of – something we intend to look into in the future – to have an in-depth discussion on domain-specific predictions and effective ways of exploiting domain information for the benefit of annotation quality and robustness.