Table 3: Dataset Statistics. Please note that in Q2B-3M dataset, character 3-grams and 4-grams tokens were also included in the vocabulary for Astec, Parabel *etc.*

Dataset	Train Instances N	Features V	Labels L	Number of Test Instances	Average Labels per sample	Average Points per label	Average Features per instance
WikiSeeAlsoTitles-350K	629,418	91,414	352,072	162,491	2.33	5.24	2.73
WikiTitles-500K	1,699,722	185,479	501,070	722,678	4.89	23.62	2.73
AmazonTitles-670K	485,176	66,666	670,091	150,875	5.39	5.11	5.26
AmazonTitles-3M	1,517,620	165,431	2,812,281	655,479	35.06	27.09	7.58
Q2B-3M	21,561,529	1,284,191	3,192,113	6,995,038	-	-	-

Table 4: Accuracy gain (Δ P@1) and training speedup for leading methods in DeepXML framework relative to the original algorithms on the AmazonTitles-670K dataset.

Method	Δ P@1	Speed-up
DeepXML + XML-CNN	+1.89	10×
DeepXML + MACH	+2.41	5×

A SUPPLEMENTARY MATERIAL

A.1 Hyper-parameters

Astec's hyper-parameters include α for combining the classifier & shortlist scores, $\hat{L}$ which was the number of labels in the surrogate task which was set to 2^{16} across all datasets and the label shortlist size which was set to 500 in all cases. Note that the different values of α leads to a trade-off in vanilla and propensity scored metrics. The embeddings in $\mathcal{Z}$, residual matrices and classifiers were initialized with FastText [23], the identity matrix and Xavier's method respectively. Astec was trained using the Adam optimizer with spectral norm constraints [38] and its hyper-parameters included the learning rate, the batch size and the number of epochs. Most of these were set to default values across datasets and the most expensive hyper-parameter to tune was the learning rate on the surrogate task.

Experiments were performed on a P40 GPU card with CUDA 11, and Pytorch 1.8 unless stated otherwise. Dropout with probability 0.5 was used for all datasets. HNSW [31] parameters M, efC and efS where set to $100, 300, 300$ and $50, 50, 100$ for ANNS^{x} and ANNS$^{\mu}$ respectively. The surrogate learning task was trained with $|\hat{L}| = 2^{16}$ with learning rate chosen from $\{0.003, 0.005, 0.02\}$. Increasing $|\hat{L}|$ beyond 2^{16} lead to only marginal gains in accuracy but at the cost of increase in training time. The model parameters for the extreme task were learnt with a learning rate chosen from $\{0.002, 0.0005\}$. It should be noted that no hyper-parameter tuning was done for proprietary datasets where Astec lead to significant gains in offline as well as online metrics.

Table 5: Parameter settings for Astec on different datasets.

Dataset	$\vert\hat{\text{L}}\vert$	Learning Rate (Surrogate task)	Learning Rate (Extreme task)
WikiSeeAlsoTitles-350K	2^{16}	0.005	0.002
WikiTitles-500K	2^{16}	0.005	0.0005
AmazonTitles-670K	2^{16}	0.02	0.002
AmazonTitles-3M	2^{16}	0.003	0.0005
Q2B-3M	2^{16}	0.02	0.002

A.2 Clustering labels

Astec clustered the labels using label centroid representation $\hat{\boldsymbol{\mu}}_l = \frac{\sum_{i=1}^{N} y_{il}\mathbf{x}_i}{\left\|\sum_{i=1}^{N} y_{il}\mathbf{x}_i\right\|_2}$ as the label meta-data was unavailable. 2-means++ algorithm was deployed to solve the following optimization problem, which recursively clusters the labels into two balanced-partitions to finally end up with $\hat{L}$ clusters.

Table 6: Astec's results for different choices of modules. Note that results are reported without re-ranker component and only one component was varied at a time. AmazonTitles-670K & WikiTitles-500K were used for (a)–(c) & (d) respectively.

(a) Intermediate representation

Method	**P@1**	**P@5**
CNN [28]	36.91	30.76
MLP [32]	37.41	30.65
Bert [13]	36.15	29.65
Astec	39.12	32.07

(b) Surrogate task

Method	**P@1**	**P@5**
Unsupervised [23]	34.80	27.42
Random [32]	38.11	30.97
Label selection	38.3	31.25
Astec	39.12	32.07

(c) Classifier

Classifier	**P@1**	**P@5**
Parabel	37.07	28.75
Slice	36.73	30.07
DiSMEC	38.42	31.44
Astec	39.12	32.07

(d) Negative sampling

Method	**P@1**	**P@5**
Uniform	27.3	11.4
NEG [34]	30.62	11.87
Slice [19]	32.33	14.37
Astec	45.45	17.64

$$
\begin{aligned}
\arg\min_{\boldsymbol{\mu}_{\pm},\boldsymbol{\alpha}\in\{-1,1\}^L} & \sum_{l=1}^{L} C_{ll}\left(\frac{1+\boldsymbol{\alpha}_l}{2}\hat{\mathbf{u}}_l\boldsymbol{\mu}_{+} + \frac{1-\boldsymbol{\alpha}_l}{2}\hat{\mathbf{u}}_l\boldsymbol{\mu}_{-}\right) \\
& + \sum_{l=1}^{L}\sum_{p=1}^{L} C_{lp}\left(\frac{1+\boldsymbol{\alpha}_l}{2}\hat{\mathbf{u}}_p\boldsymbol{\mu}_{+} + \frac{1-\boldsymbol{\alpha}_l}{2}\hat{\boldsymbol{\mu}}_p\boldsymbol{\mu}_{-}\right) \\
\text{s. t.} \quad & \|\boldsymbol{\mu}_{\pm}\|_2 = 1, \; -1 \leq \sum_{l=1}^{L}\boldsymbol{\alpha}_l \leq 1
\end{aligned}
$$

where it has been assumed without loss of generality that L labels need to be partitioned at each step, $\alpha_l = \pm 1$ means that label l is assigned to cluster with mean $\boldsymbol{\mu}_{\pm}$ and C is the $\ell 1$ normalized label correlation matrix ($\mathbf{Y}^{\top}\mathbf{Y}$). In practise, the label correlation matrix was estimated by performing a random walk over a graph with labels as nodes and data-points as edges.

A.3 Additional surrogate tasks

A simple but effective way to train the parameters of chosen feature architecture, *i.e.,* $\mathcal{Z}$ could be to select $\hat{L}$ labels based on label frequency in the training set. The hyper-parameter $\hat{L}$ was chosen to balance two constraints. First, $\hat{L}$ should be large enough so that almost all the token embeddings could be learnt in this first phase of training. At the same time, the $|\hat{L}|$ should be small enough so that (1) could be optimized efficiently on a single GPU as a non-extreme problem and without resorting to ANNS shortlisting. It was empirically observed that setting $0.05L \leq \hat{L} \leq 0.2L$ with lower values being preferred for larger problems resulted in accurate intermediate representations. It should be reiterated that the balanced clustering was found to be more accurate and scalable than the alternatives including label selection based techniques.

A.4 ANNS search and multiple representatives

It is worth pointing out that some of the most frequently occurring head labels could be usefully represented by *multiple* k-means cluster centres while constructing the ANNS small world graph. This allowed for the accurate shortlisting of multi-modal head labels which could not be shortlisted well based on a single label centroid representation, *i.e.,* $\boldsymbol{\mu}_l^0$. For instance, representing the top 4 head labels on the WikiTitles-500K dataset by 300 k-means cluster centres rather than just the label mean improved recall@300 by 5% without any noticeable increase in the training or prediction time.

A.5 Evaluation metrics

Performance has been evaluated using propensity scored precision@k and nDCG@k, which are unbiased and more suitable metric in the extreme multi-labels setting [3, 20, 40, 41]. The propensity model and values available on The Extreme Classification Repository [6] were used. Performance has also been evaluated using vanilla precision@k and nDCG@k (with k = 1, 3 and 5) for extreme classification.

For a predicted score vector $\hat{\mathbf{y}} \in R^L$ and ground truth vector $\mathbf{y} \in \{0,1\}^L$:

$$
P@k = \frac{1}{k}\sum_{l\in rank_k(\hat{\mathbf{y}})} y_l
$$

$$PSP@k = \frac{1}{k} \sum_{l \in rank_k(\hat{\mathbf{y}})} \frac{y_l}{p_l}$$

$$DCG@k = \frac{1}{k} \sum_{l \in rank_k(\hat{\mathbf{y}})} \frac{y_l}{\log(l+1)}$$

$$PSDCG@k = \frac{1}{k} \sum_{l \in rank_k(\hat{\mathbf{y}})} \frac{y_l}{p_l \log(l+1)}$$

$$nDCG@k = \frac{DCG@k}{\sum_{l=1}^{\min(k, ||\mathbf{y}||_0)} \frac{1}{\log(l+1)}}$$

$$PSnDCG@k = \frac{PSDCG@k}{\sum_{l=1}^{k} \frac{1}{\log l+1}}$$

Here, p_l is propensity score of the label l proposed in [20].

A.6 Theorem proofs

Proof. **(Bound on $||\hat{\mathbf{x}}_i - \mathbf{v}_i||_2$).** For notational convenience, we use $\hat{\mathbf{x}} := \hat{\mathbf{x}}_i$ and $\mathbf{v} := \mathbf{v}_i$;

$$\hat{\mathbf{x}}_i - \mathbf{v}_i = ReLU(\mathbf{R}\mathbf{v})$$
$$||\hat{\mathbf{x}}_i - \mathbf{v}_i||_2 = ||ReLU(\mathbf{R}\mathbf{v})||_2$$

$$\text{Using, } ||ReLU(\mathbf{u})||_2 \leq ||\mathbf{u}||_2$$
$$\leq ||\mathbf{R}\mathbf{v}||_2$$

$$\text{Using, } ||\mathbf{R}||_{op} \leq \lambda$$
$$\leq \lambda ||\mathbf{v}||_2$$

(Bound on $\hat{\mathbf{x}}_i$).

$$||\hat{\mathbf{x}}_i||_2 \leq ||\mathbf{v}||_2 + ||\hat{\mathbf{x}}_i - \mathbf{v}_i||_2$$

$$\text{Using, } ||\hat{\mathbf{x}}_i - \mathbf{v}_i||_2 \leq \lambda ||\mathbf{v}||_2$$
$$= ||\mathbf{v}||_2 + \lambda ||\mathbf{v}||_2$$
$$= (1+\lambda) ||\mathbf{v}||_2$$

In order to prove bound on cosine similarity, we first prove bound on $||\boldsymbol{\mu}_l - \boldsymbol{\mu}_l^0||_2$ and $||\boldsymbol{\mu}_l||_2$. For notational convenience, we use $\mathcal{P} := \mathcal{P}_l$, $\boldsymbol{\mu}^0 := \boldsymbol{\mu}_l^0$, and $\boldsymbol{\mu} := \boldsymbol{\mu}_l$,

(Bound on $\boldsymbol{\mu}_l - \boldsymbol{\mu}_l^0$).

$$\boldsymbol{\mu}_l - \boldsymbol{\mu}_l^0 = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \hat{\mathbf{x}}_i - \mathbf{v}_i$$

$$||\boldsymbol{\mu}_l - \boldsymbol{\mu}_l^0||_2 = || \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \hat{\mathbf{x}}_i - \mathbf{v}_i ||_2$$
$$= || \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} ReLU(\mathbf{R}\mathbf{v}_i) ||_2$$
$$\leq \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} ||ReLU(\mathbf{R}\mathbf{v}_i)||_2$$

$$\text{Using } ||ReLU(\mathbf{u})||_2 \leq ||\mathbf{u}||_2$$
$$\leq \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} ||\mathbf{R}\mathbf{v}_i||_2$$

$$\text{Using, } ||\mathbf{R}||_{op} \leq \lambda$$
$$\leq \frac{\lambda}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} ||\mathbf{v}_i||_2$$

$$\text{Let, V} := [||\mathbf{v}_1||_2, .., ||\mathbf{v}_{|\mathcal{P}|}||_2],$$

$$= \frac{\lambda}{|\mathcal{P}|}||\text{V}||_1$$

$$\text{As, } ||\text{V}||_1 \leq \sqrt{|\mathcal{P}|}||\text{V}||_2$$

$$\leq \frac{\lambda}{\sqrt{|\mathcal{P}|}}||\text{V}||_2$$

$$\leq \frac{\lambda}{\sqrt{|\mathcal{P}|}}[\sum_{i\in\mathcal{P}} ||\mathbf{v}_i||_2^2]^{\frac{1}{2}}$$

$$= \frac{\lambda}{\sqrt{|\mathcal{P}|}}[\sum_{i\in\mathcal{P}} \mathbf{v}_i{}^T\mathbf{v}_i]^{\frac{1}{2}}$$

$$\text{As, } \textstyle\sum_{i\in\mathcal{P}} \mathbf{v}_i^T\mathbf{v}_i \leq [\sum_{i\in\mathcal{P}} \mathbf{v}_i]^T[\sum_{i\in\mathcal{P}} \mathbf{v}_i]$$

$$\leq \frac{\lambda}{\sqrt{|\mathcal{P}|}}\{[\sum_{i\in\mathcal{P}} \mathbf{z}_i]^T[\sum_{i\in\mathcal{P}} \mathbf{z}_i]\}^{\frac{1}{2}}$$

$$\leq \frac{\lambda|\mathcal{P}|}{\sqrt{|\mathcal{P}|}}[\boldsymbol{\mu}^{0\top}\boldsymbol{\mu}^0]^{\frac{1}{2}}$$

$$= \lambda\sqrt{|\mathcal{P}|}||\boldsymbol{\mu}^0||_2$$

(Bound on $\boldsymbol{\mu}_l$).

$$\boldsymbol{\mu} = \frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} \hat{\mathbf{x}}_i$$

$$||\boldsymbol{\mu}||_2 = ||\frac{1}{|\mathcal{P}|}\sum_{i\in\mathbf{P}} \hat{\mathbf{x}}_i||_2$$

$$= ||\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} \mathbf{v}_i + \frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} \hat{\mathbf{x}}_i - \mathbf{v}_i||_2$$

$$\leq ||\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} \mathbf{v}_i||_2 + ||\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} \hat{\mathbf{x}}_i - \mathbf{v}_i||_2$$

$$= ||\boldsymbol{\mu}^0||_2 + ||\boldsymbol{\mu}_l - \boldsymbol{\mu}_l^0||_2$$

$$\leq (1 + \lambda\sqrt{|\mathcal{P}|})||\boldsymbol{\mu}^0||_2$$

(Lower bound on $C(\hat{\mathbf{x}}, \boldsymbol{\mu}_l)$).

$$\frac{C(\hat{\mathbf{x}}, \boldsymbol{\mu})}{C(\mathbf{v}, \boldsymbol{\mu}^0)} = \frac{\hat{\mathbf{x}}^\top\boldsymbol{\mu} \cdot ||\mathbf{v}||_2||\boldsymbol{\mu}^0||_2}{\mathbf{v}^\top\boldsymbol{\mu}^0 \cdot ||\hat{\mathbf{x}}||_2||\boldsymbol{\mu}||_2}$$

$$\text{Using, } ||\hat{\mathbf{x}}||_2 \leq (1+\lambda)||\mathbf{v}||_2 \text{ and, } ||\boldsymbol{\mu}||_2 \leq (1 + \lambda\sqrt{|\mathcal{P}|})||\boldsymbol{\mu}^0||_2$$

$$\geq \frac{\hat{\mathbf{x}}^\top\boldsymbol{\mu}}{(1+\lambda)(1+\lambda\sqrt{|\mathcal{P}|}) \cdot \mathbf{v}^\top\boldsymbol{\mu}^0}$$

$$\geq \frac{(\hat{\mathbf{x}})^\top(\boldsymbol{\mu})}{(1+\lambda\sqrt{|\mathcal{P}|})^2 \cdot \mathbf{v}^\top\boldsymbol{\mu}^0}$$

$$\text{Using, } \hat{\mathbf{x}}^\top\boldsymbol{\mu} \geq \mathbf{v}^\top\boldsymbol{\mu}^0\text{, as, } \hat{\mathbf{x}} - \mathbf{v} \geq 0\text{, and } \boldsymbol{\mu} - \boldsymbol{\mu}^0 \geq 0$$

$$\geq \frac{1}{(1+\lambda\sqrt{|\mathcal{P}|})^2}$$

(Upper bound on $C(\hat{\mathbf{x}}_i, \boldsymbol{\mu}_l)$)

$$C(\hat{\mathbf{x}}, \boldsymbol{\mu}_l) = \frac{\hat{\mathbf{x}}^\top\boldsymbol{\mu}}{||\hat{\mathbf{x}}||_2||\boldsymbol{\mu}||_2}$$

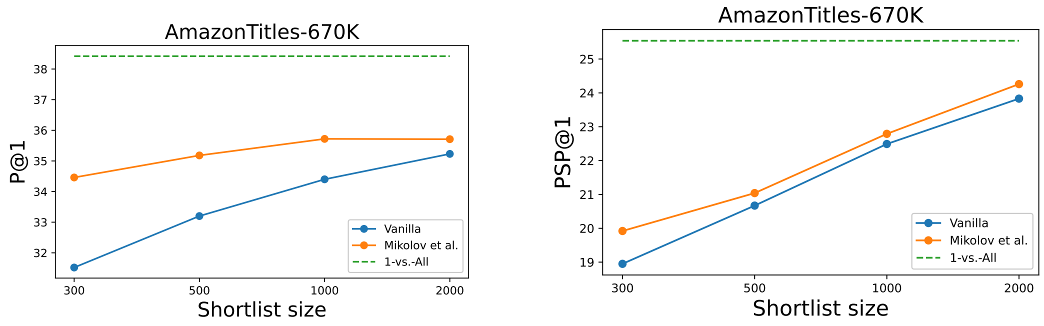

Figure 2: Performance of DeepXML when trained with a shortlist of randomly sampled negatives as compared to 1-vs.-All strategy. Vanilla strategy samples labels uniformly at random whereas Mikolov et al. samples label based on a unigram distribution over label frequencies. Astec's architecture was used for these experiments

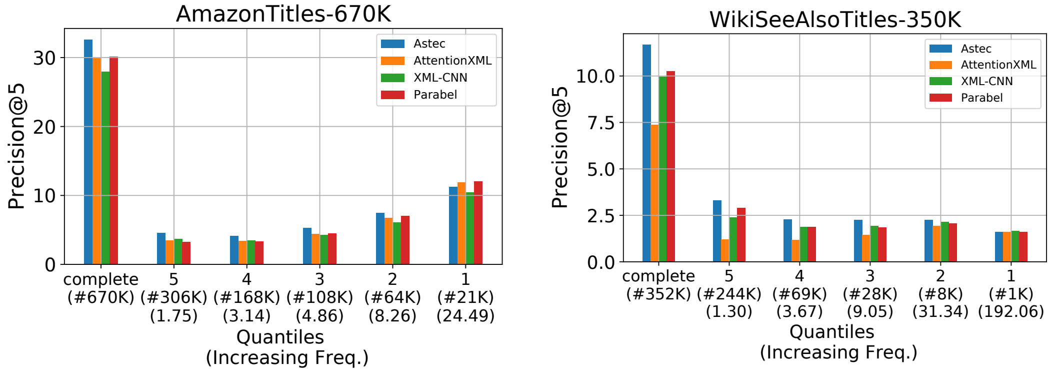

Figure 3: Quantile analysis of gains offered by Astec in terms of contribution to P@5 on the WikiSeeAlsoTitles-350K and AmazonTitles-670K datasets. The label set was divided into five equal sized bins (mean frequency in parenthesis). Astec gains are more prominent on data-scarce tail labels

Using, $||\hat{\mathbf{x}}||_2 \geq ||\mathbf{v}||_2$ and, $||\boldsymbol{\mu}||_2 \geq ||\boldsymbol{\mu}^0||_2$

$$
\begin{aligned}
&\leq \frac{\hat{\mathbf{x}}^\top \boldsymbol{\mu}}{||\mathbf{v}||_2 ||\boldsymbol{\mu}^0||_2} \\
&= \frac{\mathbf{v}^\top \boldsymbol{\mu}^0 + \mathbf{v}^\top (\boldsymbol{\mu} - \boldsymbol{\mu}^0) + (\hat{\mathbf{x}} - \mathbf{v})^\top \boldsymbol{\mu}^0 + (\hat{\mathbf{x}} - \mathbf{v})^\top (\boldsymbol{\mu} - \boldsymbol{\mu}^0)}{||\mathbf{v}||_2 ||\boldsymbol{\mu}^0||_2} \\
&= C(\mathbf{v}, \boldsymbol{\mu}^0) \\
&+ \frac{||\mathbf{v}^\top \boldsymbol{\mu}||_2 + ||\mathbf{v}^\top \boldsymbol{\mu}^0||_2 + ||\hat{\mathbf{x}}^\top \boldsymbol{\mu}||_2 - ||\mathbf{v}^\top \boldsymbol{\mu}||}{||\mathbf{v}||_2 ||\boldsymbol{\mu}^0||_2} \\
&\leq C(\mathbf{v}, \boldsymbol{\mu}^0) \\
&+ \frac{||\mathbf{v}||_2 ||\boldsymbol{\mu}||_2 + ||\mathbf{v}||_2 ||\boldsymbol{\mu}^0||_2 + ||\hat{\mathbf{x}}||_2 ||\boldsymbol{\mu}||_2 - ||\mathbf{v}||_2 ||\boldsymbol{\mu}||}{||\mathbf{v}||_2 ||\boldsymbol{\mu}^0||_2}
\end{aligned}
$$

Using, $||\hat{\mathbf{x}}||_2 \leq (1 + \lambda)||\mathbf{v}||_2$ and $||\boldsymbol{\mu}||_2 \leq (1 + \lambda\sqrt{|\mathcal{P}|})||\boldsymbol{\mu}^0||_2$

$$
\begin{aligned}
&\leq C(\mathbf{v}, \boldsymbol{\mu}^0) \\
&+ \frac{((1 + \lambda)(1 + \lambda\sqrt{|\mathcal{P}|}) - 1)||\mathbf{v}||_2 ||\boldsymbol{\mu}^0||_2}{||\mathbf{v}||_2 ||\boldsymbol{\mu}^0||_2} \\
&\leq C(\mathbf{v}, \boldsymbol{\mu}^0) + (1 + \lambda\sqrt{|\mathcal{P}|})^2 - 1
\end{aligned}
$$

□

Table 7: Astec's predicted tags for the Wikipedia title "Confederate Secret Service" are more accurate and diverse as compared to leading methods. All mispredictions have been *italicized*.

Method	Predictions
Astec	1865 disestablishments in the Confederate States of America, Government of the Confederate States of America 1861 establishments in the Confederate States of America, *Economic history of the Confederate States of America*, Military history of the Confederate States of America
Astec (without re-ranker)	1865 disestablishments in the Confederate States of America, Government of the Confederate States of America 1861 establishments in the Confederate States of America, *Economic history of the Confederate States of America* *Confederate States of America monuments and memorials*
XML-CNN	1865 disestablishments in the Confederate States of America, 1861 establishments in the Confederate States of America *American films*, *English-language films*, *Black-and-white films*
AttentionXML	*American films*, *English-language films*, Military history of the Confederate States of America, *2011 television episodes* *English-language television programming*

Table 8: Astec's predicted ads for the user Webpage title & URL Masking Tapes products - "Grainger Industrial Supply & https://www.grainger.com/search/adhesives-sealants-and-tape/tapes/masking-tapes" are more accurate and diverse as compared to leading methods (M1–M2) in Bing.

Method	Predictions
Astec	cheap masking tape, automotive paint masking tape, masking tape in bulk, blue masking tape black masking tape, 3 inch masking tape
M1	3m reflective tape, safety tape, 3m tape products, 3m packing tape, 3m 250 tape
M2	online industrial supply, industrial supply company, industrial supply inc, industrial supply national industrial supply company

Table 9: Results on AmazonImagesCat-13K dataset

Method	P@1	P@3	P@5	N@3	N@5
DeepXML	**77.19**	**54.8**	41.45	**63.48**	**59.25**
MACH	73.57	53.88	**41.99**	61.8	58.76
Slice	48.31	35.79	27.75	41.22	39.71
DiSMEC	65.69	46.63	36.76	53.82	51.62
Parabel	64.13	45.7	36.00	52.88	50.7

Table 10: Results on full-text datasets. Please note that '*' marked algorithms uses slightly different version of the dataset. Values indicated by '-' were not available.

Method	P@1	P@3	P@5	N@1	N@3	N@5	PSP@1	PSP@3	PSP@5	PSN@1	PSN@3	PSN@5
Wikipedia-500K												
Astec	71.68	50.73	39.39	71.67	62.63	60.79	29.93	35.59	39.92	29.93	35.45	38.85
Astec-3	73.02	52.02	40.53	73.02	64.10	62.32	30.69	36.48	40.38	30.69	36.33	39.84
XML-CNN	59.85	39.28	29.81	59.85	48.67	46.12	-	-	-	-	-	-
XT	64.48	45.84	35.46	-	-	-	-	-	-	-	-	-
X-Transformer*	76.95	58.42	46.14	-	-	-	-	-	-	-	-	-
AttentionXML	82.73	63.75	50.41	82.73	76.56	74.86	34.00	44.32	50.15	34.00	42.99	47.69
SLICE+FastText	27.98	16.61	12.11	27.98	22.81	22.69	15.04	14.61	15.17	15.04	15.97	17.59
DiSMEC	70.20	50.60	39.70	70.20	42.10	40.50	31.20	33.40	37.00	31.20	33.70	37.10
Parabel	68.70	49.57	38.64	68.70	60.51	58.62	26.88	31.96	35.26	26.88	31.73	34.61
AnnexML	64.64	43.20	32.77	64.64	54.54	52.42	26.88	30.24	32.79	26.88	30.71	33.33
PfastreXML	59.50	40.20	30.70	59.50	30.10	28.70	29.20	27.60	27.70	29.20	28.70	28.30
ProXML	68.80	48.90	37.90	68.80	39.10	38.00	33.10	35.00	39.40	33.10	35.20	39.00
Bonsai	69.20	49.80	38.80	69.20	60.99	59.16	27.46	32.25	35.48	-	-	-
Amazon-670K												
Astec	46.37	41.54	38.03	46.37	43.97	42.53	31.30	34.23	36.92	31.30	32.95	34.18
Astec-3	47.77	42.79	39.10	47.77	45.28	43.74	32.13	35.14	37.82	32.13	33.80	35.01
XML-CNN	35.39	31.93	29.32	35.39	33.74	32.64	28.67	33.27	36.51			
XT	42.50	37.87	34.41	42.50	40.01	38.43	24.82	28.20	31.24	24.82	26.82	28.29
AttentionXML	47.58	42.61	38.92	47.58	45.07	43.50	30.29	33.85	37.13			
SLICE+FastText	33.15	29.76	26.93	33.15	31.51	30.27	20.20	22.69	24.70	20.20	21.71	22.72
DiSMEC	44.70	39.70	36.10	44.70	42.10	40.50	27.80	30.60	34.20	27.80	28.80	30.70
Parabel	44.89	39.80	36.00	44.89	42.14	40.36	25.43	29.43	32.85	25.43	28.38	30.71
AnnexML	42.39	36.89	32.98	42.39	39.07	37.04	21.56	24.78	27.66	21.56	23.38	24.76
PfastreXML	39.46	35.81	33.05	39.46	37.78	36.69	29.30	30.80	32.43	29.30	30.40	31.49
ProXML	43.50	38.70	35.30	43.50	41.10	39.70	30.80	32.80	35.10	30.80	31.70	32.60

Table 11: Astec could be significantly more accurate and scalable than leading deep extreme classifiers including MACH, XML-CNN and AttentionXML on publicly available short-text benchmark datasets.

Method	P@1	P@3	P@5	N@1	N@3	N@5	PSP@1	PSP@3	PSP@5	PSN@1	PSN@3	PSN@5
WikiSeeAlsoTitles-320K												
Astec	20.42	14.44	11.39	20.42	19.90	20.63	9.83	12.05	13.94	9.83	11.67	12.90
Astec-3	20.61	14.58	11.49	20.61	20.08	20.80	9.91	12.16	14.04	9.91	11.76	12.98
MACH	14.79	9.57	7.13	14.79	13.83	14.05	6.45	7.02	7.54	6.45	7.20	7.73
XML-CNN	17.75	12.34	9.73	17.75	16.93	17.48	8.24	9.72	11.15	8.24	9.40	10.31
XT	16.55	11.37	8.93	16.55	15.88	16.47	7.38	8.75	10.05	7.38	8.57	9.46
SLICE+fastText	18.13	12.87	10.29	18.13	17.71	18.52	8.63	10.78	12.74	8.63	10.37	11.63
AttentionXML	15.86	10.43	8.01	15.86	14.59	14.86	6.39	7.20	8.15	6.39	7.05	7.64
DiSMEC	16.61	11.57	9.14	16.61	16.09	16.72	7.48	9.19	10.74	7.48	8.95	9.99
Parabel	17.24	11.61	8.92	17.24	16.31	16.67	7.56	8.83	9.96	7.56	8.68	9.45
AnnexML	14.96	10.20	8.11	14.96	14.20	14.76	5.63	7.04	8.59	5.63	6.79	7.76
PfastreXML	15.09	10.49	8.24	15.09	14.98	15.59	9.03	9.69	10.64	9.03	9.82	10.52
Bonsai	17.95	12.27	9.56	17.95	17.13	17.66	8.16	9.68	11.07	8.16	9.49	10.43
AmazonTitles-670K												
Astec	39.97	35.73	32.59	39.97	37.91	36.60	27.59	29.79	31.71	27.59	28.80	29.61
Astec-3	40.63	36.22	33.00	40.63	38.45	37.09	28.07	30.17	32.07	28.07	29.20	29.98
MACH	34.92	31.18	28.56	34.92	33.07	31.97	20.56	23.14	25.79	20.56	22.18	23.53
XML-CNN	35.02	31.37	28.45	35.02	33.24	31.94	21.99	24.93	26.84	21.99	23.83	24.67
XT	36.57	32.73	29.79	36.57	34.64	33.35	22.11	24.81	27.18	22.11	23.73	24.87
SLICE+fastText	33.85	30.07	26.97	33.85	31.97	30.56	21.91	24.15	25.81	21.91	23.26	24.03
AttentionXML	37.92	33.73	30.57	37.92	35.78	34.35	24.24	26.43	28.39	24.24	25.48	26.33
DiSMEC	38.12	34.03	31.15	38.12	36.07	34.88	22.26	25.46	28.67	22.26	24.30	26.00
Parabel	38.00	33.54	30.10	38.00	35.62	33.98	23.10	25.57	27.61	23.10	24.55	25.48
AnnexML	35.31	30.90	27.83	35.31	32.76	31.26	17.94	20.69	23.30	17.94	19.57	20.88
PfastreXML	32.88	30.54	28.80	32.88	32.20	31.85	26.61	27.79	29.22	26.61	27.10	27.59
Bonsai	38.46	33.91	30.53	38.46	36.05	34.48	23.62	26.19	28.41	23.62	25.16	26.21
WikiTitles-500K												
Astec	46.01	25.62	18.18	46.01	34.58	32.82	18.62	18.59	18.95	18.62	20.01	21.64
Astec-3	46.60	26.03	18.50	46.60	35.10	33.34	18.89	18.90	19.30	18.89	20.33	22.00
MACH	33.74	15.62	10.41	33.74	22.61	20.80	11.43	8.98	8.35	11.43	10.77	11.28
XML-CNN	43.45	23.24	16.53	43.45	31.69	29.95	15.64	14.74	14.98	15.64	16.17	17.45
XT	39.44	21.57	15.31	39.44	29.17	27.65	15.23	15.00	15.25	15.23	16.23	17.59
SLICE+fastText	28.07	16.78	12.28	28.07	22.97	22.87	15.10	14.69	15.33	15.10	16.02	17.67
AttentionXML	42.89	22.71	15.89	42.89	30.92	28.93	15.12	14.32	14.22	15.12	15.69	16.75
DiSMEC	39.89	21.23	14.96	39.89	28.97	27.32	15.89	15.15	15.43	15.89	16.52	17.86
Parabel	42.50	23.04	16.21	42.50	31.24	29.45	16.55	16.12	16.16	16.55	17.49	18.77
AnnexML	39.56	20.50	14.32	39.56	28.28	26.54	15.44	13.83	13.79	15.44	15.49	16.58
PfastreXML	30.99	18.07	13.09	30.99	24.54	23.88	17.87	15.40	15.15	17.87	17.38	18.46
Bonsai	42.60	23.08	16.25	42.60	31.34	29.58	17.38	16.85	16.90	17.38	18.28	19.62
AmazonTitles-3M												
Astec	47.64	44.66	42.36	47.64	45.89	44.66	15.88	18.59	20.60	15.88	17.71	19.02
Astec-3	48.74	45.70	43.31	48.74	46.96	45.67	16.10	18.89	20.94	16.10	18.00	19.33
MACH	37.10	33.57	31.33	37.10	34.67	33.17	7.51	8.61	9.46	7.51	8.23	8.76
XT	27.99	25.24	23.57	27.99	25.98	24.78	4.45	5.06	5.57	4.45	4.78	5.03
SLICE+fastText	35.39	33.33	31.74	35.39	34.12	33.21	11.32	13.37	14.94	11.32	12.65	13.61
AttentionXML	46	42.81	40.59	46.00	43.94	42.61	12.81	15.03	16.71	12.80	14.23	15.25
Parabel	46.42	43.81	41.71	46.42	44.86	43.70	12.94	15.58	17.55	12.94	14.70	15.94
AnnexML	48.37	44.68	42.24	48.37	45.93	44.43	11.47	13.84	15.72	11.47	13.02	14.15
PfastreXML	31.16	31.35	31.10	31.16	31.78	32.08	22.37	24.59	26.16	22.37	23.72	24.65
Bonsai	46.89	44.38	42.30	46.89	45.46	44.35	13.78	16.66	18.75	13.78	15.75	17.10