BAN: Large Scale Brand ANonymization for Creative Recommendation via Label Light Adaptation

Keqian Li[†], Kevin Yen[†], Shaunak Misra[†], Yifan Hu[†], Changwei Hu[¶], Manisha Verma[×]
Yahoo Research[†], Amazon Inc.[×], XPeng Motors[¶]

***Abstract*—One of the primary component in ads creative recommendation system is the brand anonymization that removes brand-specific information from ad text for legal compliance and providing ready-to-use template for the advertisers to customize and consume. In our previous work [1] on ads creative recommendation system, the anonymization is done via a block list created solely based on manual reviewing, which is expensive and limits in the scale of the deployment of the ads recommendation. In this work we investigate a large scale, automated approach for brand anonymization. Such a problem presents many unique and non-trivial challenges, including the domain specificity of the brand entities, the fine-granularity requirements of structured output, the tight constraint of the limited contexts, the high level of grammatical noise in the advertisement data, and the heterogeneity of information required to perform anonymization. We propose a transformer model that leverage implicit knowledge together with a label-light adaptation procedure for this task. Our model is rolled out to ads systems in Yahoo that cover billions of impression traffic per month and improved previous production system by 68.3% F1-score on token level prediction and 61.6% on ad level prediction.**

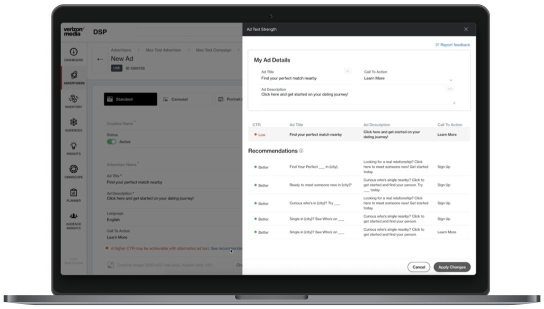
I. Introduction

When advertisers on-board to online ads platform, they are required to provide ad creative such as an attractive title and descriptive text. Coming up with effective ad creative is a time consuming process, and particularly challenging for small businesses with limited experience and resource [2]. *Ad creative recommendation* that provides advertisers with high quality creatives templates using existing high quality ads [1] are invaluable. One of the critical component in such a system is the *brand anonymization*. As shown in Figure 1, it is responsible for intelligently segmenting content by understanding brand agnostic drivers for high quality ads. With brands removed, the resulting ad template provides actionable examples for businesses create their ads for maximized audience impact.

Historically, the advertisement system follows a manual editing approach: given a set of human compiled brand names block list represented as a collection of phrases [1], an exhaustive matching in the ad text is done and any content that overlap with the block list is anonymized. This is extremely labor-intensive with a high latency and high operating costs. As the system increases its scope towards billions of ads traffic [3], this has increasingly become a critical scale bottleneck for its practical use. An automated system for ad anonymization is therefore needed.

Specifically, the task can be described as: given a set of existing high quality ad creatives and block list words, identify and segment brand entity in candidate ads as spans in the original text. In other words, we perform *concept recognition* [4] for a given result of concept extraction *concept extraction* [4], [5]. Consequentially, the ad creatives after removing brand information will be used as templates to recommend to the advertisers.

Our novel solution builds upon previous concept mining from natural language [6], [7], [8], [9] . Instead of unsupervised mining [10], [11], [12], [13], [5], crowd-sourced label are employed to generate light weight block list, and alignment into raw text. With the additional help of *knowledge fusion* incorporating structural fields of ads and advertiser registered in the database grid, large scale language models (LM) [14] are adapted to close domain concept recognition task efficiently. Evaluation shows it has significantly performance gains over previous production system, and is rolled out to both the Yahoo Gemini ads system [15], [16] and third party DSP ad exchange that covers billions of impression traffic per month.



New Ad
LINE ID 1369735

Creative Name *
Status
Active
Advertiser Name *
Ad Title *
Find your perfect match nearby
Ad Description *
Click here and get started on your dating journey!
Language
English
Call To Action
Learn More
A higher CTR may be achievable with alternative ad text. See recommenda
Creative Image (1200x627 min size). Aspect ratio: 1.9:1

My Ad Details

Ad Title
Find your perfect match nearby
Call To Action
Learn More
Ad Description
Click here and get started on your dating journey!

CTR	Ad Title	Ad Description	Call To Action
● Low	Find your perfect match nearby	Click here and get started on your dating journey!	Learn More

Recommendations

● Better	Find Your Perfect ___ in {city}.	Looking for a real relationship? Click here to meet someone new! Get started today.	Sign Up
● Better	Ready to meet someone new in {city}?	Curious who's single nearby? Click to get started and find your person. Try ___ today.	Sign Up
● Better	Curious who's in {city}? Try ___	Looking for a real relationship? Click here to meet someone new! Get started today.	Sign Up
● Better	Single in {city}? See Who's on ___	Curious who's single nearby? Click to get started and find your person.	Sign Up
● Better	Single in {city}? See Who's on ___	Curious who's single nearby? Click to get started and find your person.	Learn More

Fig. 1: BAN Production System Interface in DSP

II. The BAN framework

In production pipeline, features from Hadoop ecosystem together with raw text to label are integrated into the function, which will combine it with the block-list to generate and learn from *phrasal level label-light supervision* to produce a sequence labeling function to deploy to online ads traffic.

Problem Formulation We follow knowledge base terminology and assume the input feature set is given in the form of predicates and also assume the block list is given as raw texts. Suppose $\mathcal{A}$ is a set of ads, $\mathcal{B}$ is the set of block lists elements, $\mathcal{R}$ is a set of types of predicates, $\mathcal{L}$ is a set of raw text literals, model $\mathcal{M} \subseteq (\mathcal{A} \cup \mathcal{B}) \times \mathcal{R} \times \mathcal{L}$ is given as triplets between ads and literals for existing facts [1]. Specifically, there exists the ad text type of facts $r_{\mathcal{A}} \in \mathcal{R}$ such that

$$\forall a_i \in \mathcal{A}, \exists l_j \in \mathcal{L}, s.t., (a_i, r_{\mathcal{A}}, l_j) \in \mathcal{M}$$

and the block list text type of facts $r_{\mathcal{B}} \in \mathcal{R}$ such that

$$\forall b_i \in \mathcal{B}, \exists l_j \in \mathcal{L}, s.t.(b_i, r_{\mathcal{B}}, l_j) \in \mathcal{M},$$

We represent raw text as unsegmented character list, where $l_j = (c_{j1}, c_{j2}, c_{j3}, \ldots, c_{j|l_j|})$ for each $l_j \in \mathcal{L}$ where $|l_j|$ is the character count. The brand anonymization task is to perform *concept recognition* for brand revealing [4]. More specifically, the task can be understood as learning *sequence labeling* function $\mathbf{f} \in \mathcal{F}$ to identify the membership for each element $\{c_{ik} | a_i \in \mathcal{A}, 1 \leq k \leq |l_i|, (a_i, r_{\mathcal{A}}, l_j) \in \mathcal{M}\}$ in the ad text

Example 1[Concept Recognition via character sequence labeling] Given a input ad $a \in \mathcal{A}$ with following ad text $l \in \mathcal{L}$, such at $(A, r_{\mathcal{A}}, l) \in \mathcal{M}$, the original raw text reads

> *Huge Toyota Price Cuts + Discounts! Get July Only Toyota Specials* (Toyota Advertiser)

[1] facts and model are used interchangeably throughout the paper

978-1-6654-3902-2/21/31.002021$IEEE$

An sequence labeling function will identify the occurrence of the two *Toyota*s inside the text. As a result, a template for ads creative

Huge ___ *Price Cuts + Discounts! Get July Only* ___ *Specials*
(Candidate creative)

can be generated as one end product.

Function Mapping Due to the *heterogeneity* of input, we assume the sequence labeling function $\mathbf{f}$ take in the input knowledge base $\mathcal{M}$ and requires it to produce probabilistic output. Specifically, individual component $\mathbf{f}(\mathcal{M})_{ik}|a_i \in \mathcal{A}, 1 \leq k \leq |l_i|, (a_i, r_{\mathcal{A}}, l_j) \in \mathcal{M} \in \mathbb{R}^{\mathcal{Y}}$ indicate the predicted probability distribution over possible labels.

Text Segmentation We follow standard text segmentation function [12] $\mathbf{S}$ that given a literal $l = (c_1, c_2, c_3, \ldots, c_{|l|})$ will produce the index sequence of segment end points, $d^{(\mathbf{S})} = (s_1^{(\mathbf{S})}, s_2^{(\mathbf{S})}, \ldots, s_{|d^{(\mathbf{S})}|}^{(\mathbf{S})})$ satisfying $1 = s_1^{(\mathbf{S})} < s_2^{(\mathbf{S})} < \ldots < s_{|d^{(\mathbf{S})}|}$, where for each $1 \leq t < |d^{(\mathbf{S})}|$, there is a whole word level segment $w_t^{(\mathbf{S})} := (c_{s_t}^{(\mathbf{S})}, \ldots, c_{s_{t+1}-1}^{(\mathbf{S})})$. In our work we use a custom Spacy NLP pipeline that first preprocess the punctuation into space separated tokens and apply regex splitting.

Phrasal Level Supervision We follow a *label light* approach and generate heuristic labeling function for each $a_i \in \mathcal{A}$ and each $1 \leq k \leq |l_i|, (a_i, r_{\mathcal{A}}, l_j) \in \mathcal{M}$. The output will be positive if and only if it matches with the entire phrase in the block list. More formally, as

$$y'_{ik} := \begin{cases} 1, & \text{if } \exists t, \delta, b \in (B) \\ & s.t. (s_j)_t \leq k < (s_j)_{t+1}, \\ & ((w_j)_t \ldots (w_j)_{t+\delta}) \approx b \\ 0, & \text{otherwise} \end{cases}$$

In this work we define the approximate equal $\approx$ between two sequence as element-wise exact match after uncasing and retaining characters post regex filter.

Adaptation We exploit the fact that the literals in the knowledge base are inherently text based feature, and leverage pre-trained language model architecture $\mathbf{T}$ as the *raw text encoder* for all knowledge base information. Formally, the sequence labeling function $\mathbf{f}$ follows the neural net architecture shown below

$$\mathbf{f}_{ik} := \begin{cases} \mathbf{proj}(\mathbf{T}(\mathbf{Aug}(l_j))_t) & \text{if } \exists t, \delta, b \in (B) \\ & s.t. (s_j^{(\mathbf{S}')})_t \leq k < (s_j^{(\mathbf{S}')})_{t+1}, \\ & ((w_j^{(\mathbf{S}')})_t \ldots (w_j^{(\mathbf{S}')})_{t+\delta}) \approx b \\ \infty, & \text{otherwise} \end{cases}$$

where $proj$ is a position agnostic MLP layer for obtaining the original input, and

$$\begin{aligned} \mathbf{Aug}(l_j) := (c_{j1}, \ldots, c_{j|l_j|} || \mathbf{NL}(\{(a_i, r, l) \in \mathcal{M}\})) \\ s.t. (a_i, r_{\mathcal{A}, l_j}) \in \mathcal{M} \end{aligned} \tag{1}$$

is a concatenation of original ad text and a natural language encoding function $\mathbf{NL}$ of the triples in the knowledge base. In this work we implement the encoding function as the concatenation of original literal value.

Optimization The *BAN* model will be trained end-to-end using position wise loss

$$\arg min_{\mathbf{f} \in \mathcal{F}} \sum_{a_i \in \mathcal{A}} \sum_{1 \leq k \leq |l_i|, (a_i, r_{\mathcal{A}}, l_j) \in \mathcal{M}} \mathbf{SBCE}(\mathbf{f}_{ik}, y'_{ik})$$

where $\mathcal{A} \subseteq \mathcal{A}$ is the test set of ads, $\mathbf{SBCE}$ denotes the sparse binary cross entropy.

III. EXPERIMENT

In this section, we present a series of experiment in seek of following research questions:

RQ1 How does possible alternative methods compare to BAN along with in production environment?

RQ2 How does the possible settings in BAN affect its quality quantitatively and qualitatively?

RQ3 How much data and labeling effort is needed for the adapatation?

RQ4 How does BAN compare to baselines in computation cost? Could we achieve healthier trade-off between effectiveness and efficiency?

RQ5 How generalizable is the performance of BAN compare with alternatives in production environment in downstream application tasks?

RQ6 Is the result of BAN reliable and interpretable to human on production data?

Core Production Metric Evaluation [RQ1] We first evaluate various anonymization approaches and on the core production metric of Brand Token level Precision, Brand Token Level Recall as well as Brand Token Level F1 score. We employed an affiliated team of crowdsource workers and collected a 37584 ad text used for model adaptation, with a 1410 raw text block list. We additionally collected 9321 fully annotated ad text for evaluation. The evaluation simulates a *cold start* environment where there are no brand or advertiser overlap between the adaptation and evaluation data.

In addition to BAN, which uses a 12 layer, 12 heads, multilingual version of Bert with a vocabulary size of 120K and maximum sequence length of 256, we compare the following approaches with V0.1 block list production system.

- **Grammartical Pattern** capturing all phrases matching pre-compile POS-Tag regex [12]
- **Phrase Chunking** using pre-trained noun phrases identified by NLP model [12]
- **NER** capturing all named entities using pre-trained NER tagger with Spacy 3.0.6.
- **Brand Knowledge Base** invoking Yahoo Brand Knowledge Graph for brand name linking through a web service [17]
- **Phrase Matching**: using phrase level normalized text matching as mentioned in section II by comparing the full brand name database used by Yahoo Brand Knoweledge Graph. [17]

As shown in Table I, BAN compares favourably to production systems across all core metrics followed by production itself, with the exception of recall where rule based system have an advantage.

	Brand Token Precision		Brand Token Recall		F1 Score	
Model	Percentage	Lift	Percentage	Lift	Percentage	Lift
Grammatical Pattern	5.34%	-84.72%	89.64%	52.19%	10.08%	-77.02%
Phrase Chunking	4.72%	-86.49%	79.15%	34.38%	8.91%	-79.69%
NER	7.76%	-77.80%	52.49%	-10.88%	13.52%	-69.18%
Brand Knowledge Base	30.09%	-13.91%	12.14%	-79.39%	17.30%	-60.57%
Phrase Matching	10.14%	-70.99%	50.36%	-14.50%	16.88%	-61.52%
BAN	**80.21%**	**129.50%**	**68.43%**	**16.18%**	**73.86%**	**68.36%**
Production	34.95%	0.00%	58.90%	0.00%	43.87%	0.00%

TABLE I: Core Metric Comparison with the Production System

BAN Ablation Study [RQ2-3] We first study the impact of text pre-processing and model architecture by comparing model with different language model encoder. As shown in Figure 2, different versions of BAN tends to have very similar performance, and the differences correlates positively with complexity of the architecture and computational cost.

We then study the impact of data used for adaptation by comparing BAN with variable number of sample adaptation data. As shown in

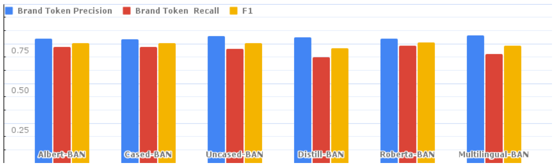


Fig. 2: Preprocessing and Architecture Ablation Evaluation

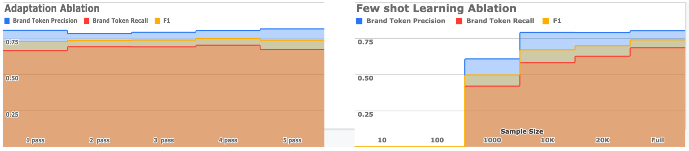


Fig. 3: Quantitative Ablation

Figure 3, as the size of data grows, the inflection point occurs around 1K training ads where performance jumps from 0 to above half and quickly recover full performance on the original data.

We Further study the adaption process by varying the number of passes the model trains. As shown in Figure 3, BAN quickly recovers performance after the 1st pass and then stabilize, showing robustness even under small amount of supervision.

computation cost [RQ4] For production system it is of premier importance to balance between computation cost and performance. Here we study the computational cost in a per mapper node setting, where shard-ed text are sent to the local machine to process sequentially. Figure 4 shows the latency in terms of seconds on a batch of 10K ads, where different versions of BAN shows advantage over approaches with large dictionary, NLP pipelines or web services, with distilled BAN variants having the lowest latency.

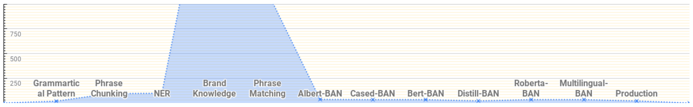


Fig. 4: Per Mapper Node Computation Cost Study

Manual evaluation [RQ6] To further validate our approach and ensure its interpret-ability and robustness, we conduct manual inspection on 200 random samples and ask domain experts to judge its performance. Review results by production team shows that generated anonymization are understandable and consistent the automated token level evaluation performed (Figure 5).

Application [RQ5] Another related task is sentence tagging, where instead of giving structured template, ads requiring are tagged and sent for further review. As shown in Figure 6, BAN performs well in term of precision, recall, F1 score 90% on many key metrics.

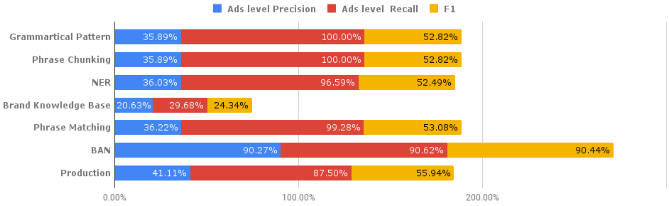
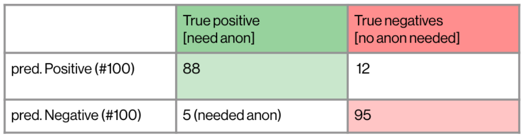


Fig. 6: Application: Ads tagging



	True positive [need anon]	True negatives [no anon needed]
pred. Positive (#100)	88	12
pred. Negative (#100)	5 (needed anon)	95

Fig. 5: Manual Evaluation Result Analysis

IV. Conclusion

In this paper we present the first automated system for brand anonymization that leverage the novel technique for label light adaptation of large pre-trained natural language models for the task of concept recognition given concept recognition. One particular interesting directions is to extend the model for general purpose concept recognition and mining tasks, another is to better integrate knowledge base to infer the identify and connections of concepts.

References

[1] S. Mishra, C. Hu, M. Verma, K. Yen, Y. Hu, and M. Sviridenko, "Tsi: an ad text strength indicator using text-to-ctr and semantic-ad-similarity," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 4036–4045.

[2] S. Mishra, M. Verma, Y. Zhou, K. Thadani, and W. Wang, "Learning to create better ads: Generation and ranking approaches for ad creative refinement," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, ser. CIKM '20, 2020.

[3] https://expandedramblings.com/index.php/google-advertising-statistics/, 2021, [Online; accessed 19-July-2008].

[4] K. Li, "Mining and analyzing technical knowledge based on concepts," Ph.D. dissertation, University of California Santa Barbara, 2019.

[5] K. Li, Y. He, and K. Ganjam, "Discovering enterprise concepts using spreadsheet tables," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1873–1882.

[6] H. Zha, J. Shen, K. Li, W. Greiff, M. T. Vanni, J. Han, and X. Yan, "Fts: Faceted taxonomy construction and search for scientific publications," 2018.

[7] K. Li, P. Zhang, H. Liu, H. Zha, and X. Yan, "Poqaa: Text mining and knowledge sharing for scientific publications," 2018.

[8] H. Zha, W. Chen, K. Li, and X. Yan, "Mining algorithm roadmap in scientific publications," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1083–1092.

[9] K. Li, S. Li, S. Yavuz, H. Zha, Y. Su, and X. Yan, "Hiercon: Hierarchical organization of technical documents based on concepts," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 379–388.

[10] K. Li, H. Zha, Y. Su, and X. Yan, "Unsupervised neural categorization for scientific publications," in *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 2018, pp. 37–45.

[11] K. Li, W. Lu, S. Bhagat, L. V. Lakshmanan, and C. Yu, "On social event organization," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1206–1215.

[12] K. Li, H. Zha, Y. Su, and X. Yan, "Concept mining via embedding," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 267–276.

[13] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[14] F. Tan, Y. Hu, C. Hu, K. Li, and K. Yen, "Tnt: Text normalization based pre-training of transformers for content moderation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4735–4741.

[15] Y. Zhou, S. Mishra, J. Gligorijevic, T. Bhatia, and N. Bhamidipati, "Understanding consumer journey using attention based recurrent neural networks," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19, 2019.

[16] N. Bhamidipati, R. Kant, and S. Mishra, "A large scale prediction engine for app install clicks and conversions," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ser. CIKM '17, 2017.

[17] M. Kejriwal, K. Shen, C.-C. Ni, and N. Torzec, "An evaluation and annotation methodology for product category matching in e-commerce," *Computers in Industry*, vol. 131, p. 103497, 2021.