Finding Lookalike Customers for E-Commerce Marketing

Yang Peng
yang.peng@walmart.com
Walmart Global Tech

Changzheng Liu
changzheng.liu@walmart.com
Walmart Global Tech

Wei Shen
wei.shen@walmart.com
Walmart Global Tech

ABSTRACT

Customer-centric marketing campaigns generate a large portion of e-commerce website traffic for Walmart. As the scale of customer data grows larger, expanding the marketing audience to reach more customers is becoming more critical for e-commerce companies to drive business growth and bring more value to customers. In this paper, we present a scalable and efficient system to expand targeted audience of marketing campaigns, which can handle hundreds of millions of customers. We use a deep learning based embedding model to represent customers and an approximate nearest neighbor search method to quickly find lookalike customers of interest. The model can deal with various business interests by constructing interpretable and meaningful customer similarity metrics. We conduct extensive experiments to demonstrate the great performance of our system and customer embedding model.

1 INTRODUCTION

In customer relationship management (CRM) systems, customer acquiring and retention are crucial for marketing success. Expanding the set of targeted customers is a very important component in CRM systems for both customer acquiring and retention. In this article, we consider the problem of building a large scale marketing audience expansion system, aiming at driving e-commerce growth by finding more customers for CRM email and push marketing campaigns.

Formally speaking, the problem to solve in this paper is: given a set of existing customers (seed customer set) for a marketing campaign, how to find more customers that are similar to these seed customers, so that we can increase revenue and drive more traffic for Walmart e-commerce.

One of the biggest challenges we face in this problem is the scale of the data. Walmart has hundreds of millions of active customers in the US market alone. Each customer could have hundreds or even thousands of features. And customer data is increasing rapidly year by year. Thus building a scalable and efficient system to handle ever-increasing big customer data is our top priority. Besides scalability, we need to take into account the interpretability of our method when finding lookalike customers. Model interpretability not only helps explain how our method works to business partners, but also provides an intuitive way for examining system quality.

In this article, we propose a scalable and efficient audience expansion system for marketing campaigns, which can handle hundreds of millions of customers.

- Our system can generate low dimensional dense embeddings to represent customers.
- In the customer embedding space, we use the cosine distance between the two customer embedding vectors as the estimation of the similarity between two customers.
- The similarity metric we design can measure different business interests (such as purchases, visits and engagements),

which are interpretable and meaningful business goals for marketing campaigns.

- We use approximate nearest neighbor search to quickly find lookalike customers to the seed customers in the customer embedding space.

To improve the quality of the customer embedding model, our model ingests multimodal features from different data sources, such as transactions, visits, engagements and customer metadata. Multimodal fusion techniques have demonstrated great benefits for tasks such as information retrieval, information extraction and classification [4, 17, 18, 20–26] by leveraging the complementary and correlative relations between different types of data. Our embedding model can also achieve better quality for audience expansion by combining different types of data.

Our contributions are shown below:

- We propose an effective and efficient marketing audience expansion system, which can handle hundreds of millions of customers.
- We use a deep learning model to generate customer embeddings. The deep learning model can handle both dense numerical features and sparse categorical features. And the model encodes location embedding using transfer learning.
- Our system has great adaptability by constructing different similarity metrics for different campaigns and business goals. The similarity metrics are both interpretable and meaningful.
- We develop a scalable and efficient approximate nearest neighbor search method based on FAISS to quickly find similar customers.
- We design multimodal features from various data sources.
- Extensive experiments have been conducted to demonstrate the scalability, efficiency and quality of our system and embedding model.

Overview Related work on lookalike modeling and audience expansion systems is discussed in Section 2. The overview of our system is presented in Section 3. The embedding model is illustrated in Section 4. We demonstrate the effectiveness and efficiency of our system through extensive experiments in Section 5. The conclusions and future work of our lookalike model are discussed in Section 6.

2 RELATED WORK

In this section, we will discuss the previous work on audience expansion systems, use representation models and fast similarity search methods. For the literature review, we mostly focus on related work in industry solutions, since we are tacking large-scale or even web-scale datasets which are very rarely studied in academic research.
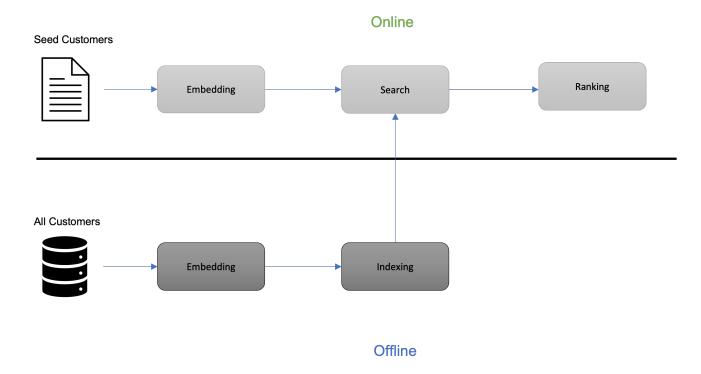
Figure 1: Audience Expansion System in Walmart.

2.1 Audience Expansion

In [16], Ma et.al. discussed three types of approaches for audience expansion for marketing campaigns: similarity-based, regression-based and segmentation-based approaches. They proposed a graph-based lookalike system in Yahoo! advertising platform, which takes advantages of both simple similarity and regression-based methods. In [14], Liu et.al. developed two methods to achieve audience expansion in LinkedIn: campaign-agnostic expansion based on user attributes and campaign-aware expansion using nearest neighbor search. In [8], Jiang et.al. discussed rule-based, similarity-based and model-based methods for finding lookalike users and proposed a deep neural network classification model for audience expansion in MiningLamp. In [3], deWet et.al. proposed a two-stage embedding-based audience expansion model that is deployed in production at Pinterest. For the first stage, they trained a global user embedding model on sitewide user activity logs. In the second stage, they used statistical techniques to create lightweight seed list representations in the embedding space for each advertiser.

In our system, we first use a similarity-based approach to search lookalike customers from the whole customer universe and then rank these new customers based on their similarity scores or a separate classification model. We choose the similarity-based approach for the first step because of its great scalability and low search latency. Similarity-based approaches usually require building user representations first. In the next section, we will discuss embedding models for representing customers.

2.2 User Representation

Embedding models are very useful in terms of transforming high dimensional sparse feature vectors of customers to low dimensional dense representations of customers. Embedding models have been widely used in industry, for example, for search engine marketing at Walmart [9–11], search ranking at Airbnb [5], and recommendation at Pinterest [19]. Embedding models can be trained in a way to capture similarity between customers, so that we can use approximate nearest neighbor search methods to quickly find lookalike customers of seed customers. In our case, we use a two-tower architecture to train the customer embedding model, which is well recognized in previous work. Our novelty in user representation is modeling business metrics using the cosine similarity between two embedding vectors, which has great interpretability.

2.3 Similarity Search

After user representation, we need a fast approach to search looka-like customers from a very large customer universe with potential size of hundreds of millions of customers. Scanning the whole customer universe is not scalable or efficient [1, 15, 27–29]. Approximate nearest neighbor search is the most popular approach used in previous work. For example, locality sensitive hashing (LSH) has been used in previous work [3, 14, 16]. There are several good open-source tools [7, 13, 30–34] for approximate nearest neighbor search, such as ScaNN [6] by Google and FAISS [12] by Facebook. We choose FAISS for lookalike customer search for its scalability of handling billions of vectors and support of various types of distance measures (e.g. dot-product, cosine, Euclidean distances).

3 SYSTEM OVERVIEW

In this section, we explain our audience expansion system pipeline for finding lookalike customers for e-commerce marketing in Walmart. The system diagram is shown in Figure 1. Our audience expansion system has an online stage and an offline stage. In the offline stage, we generate customer embeddings for all the customers in the customer universe and then build indexing on the customer embedding space. In the online stage, the seed customers are transformed into embeddings and then we search for lookalike customers using the pre-built indexes from the offline stage. After getting the lookalike customers, we will conduct filtering and ranking on them. The ranking approach can be based on their similarity scores or a different classification model.

3.1 Embedding

In this article, we will mostly focus on the embedding model and explain how we build this model in later sections. The customer embedding model yields unified dense representations of customers. Our customer embedding model can be utilized in a lot of use cases. Besides finding lookalike customers, customer embeddings can be employed as input features in other customer models, such as purchase propensity models and life-time value models. Ranking method is not a focus in this paper and will be studied in our future work.

3.2 Indexing and Search

The scalability issue in this system is how to quickly search for lookalike customers from a pool of hundreds of millions of candidate customers. To narrow down the customers for consideration, we build indexes using FAISS [12]. We choose FAISS for a few reasons as listed below. FAISS supports a wide range of different indexes and provides both CPU and GPU implementations. FAISS also supports different types of distance measures, such as Euclidean distances, dot products and cosine similarity. And we can utilize compression techniques in FAISS to process large datasets that cannot fit in memory. How we implement indexing and search is not the main focus of this paper. If you are interested in more details about FAISS, please visit their Github project page.

4 EMBEDDING MODEL

In this section, we present our deep learning based embedding model, which can capture the similarity between customers. To design this model, we need to first define the similarity metric between two customers. While some models in previous work [3] learned relative similarity scores between positive and negative customer pairs, we use direct similarity metrics between two customers, which are interpretable, meaningful and reflecting business metrics. The similarity metrics we define can be very useful in improving and explaining marketing campaign performance.

We use a two-tower architecture to calculate the cosine distance between two customer embedding vectors and use the cosine distance as the estimate of similarity metric between two customers. The customer embedding model is trained to minimize the total loss between cosine distances and true similarity scores in the training datasets.

4.1 Similarity Metric Definition

In Walmart, we care about a lot of different business metrics for marketing campaigns, such as transactions, website visits, campaign engagements. When expanding audience for marketing campaigns, it's a business advantage to find new customers that have similar behavior on Walmart e-commerce website as the existing seed customers. Finding new customers with similar business metrics can allow marketing campaigns to maintain a similar customer distribution after expansion, which is very beneficial for cold-start campaigns or conversion campaigns.

There are three types of business metrics of particular interest to our marketing campaigns: transactions, visits and engagements. Let's take transactions as an example to illustrate how we define the similarity metrics. Let's say there are a list of product categories in Walmart catalog, $c_1, c_2, ..., c_n$. Customer A has made purchase orders in these categories, $O_A = (a_1, a_2, ..., a_n)$. Customer B has also made purchase orders in these categories, $O_B = (b_1, b_2, ..., b_n)$. The similarity between A and B is defined as:

$$similarity(A, B) = cosine\_similarity(O_A, O_B) \qquad (1)$$

We can also use other similarity distance functions, such as Jaccard similarity and Euclidean distance. This method of calculating similarity metrics can also be applied for visits and engagements.
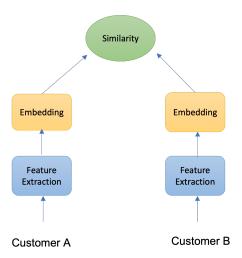

Figure 2: Two-Tower Model Architecture.

4.2 Two-Tower Architecture

After defining the similarity metric between customers, the next task is to build a machine learning model to predict the similarity metric given customer features. Our approach is:

(1) extract raw features for customers;
(2) transform customer features into customer embeddings;
(3) calculate the cosine distances of customer embedding pairs in the embedding space;
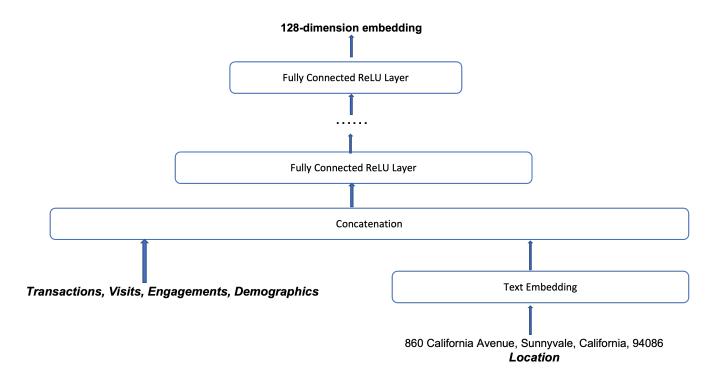(4) use the cosine distances as the estimates of true similarity scores of customer pairs.

128-dimension embedding

Fully Connected ReLU Layer

· · · · · ·

Fully Connected ReLU Layer

Concatenation

Transactions, Visits, Engagements, Demographics

Text Embedding

860 California Avenue, Sunnyvale, California, 94086
Location

Figure 3: The Embedding Model.

The process is shown in Figure 2. The loss function for optimization is the L1 loss between cosine distance prediction and true similarity metric.

The multimodal customer features are extracted from multiple data sources, including transaction data, visit data, engagement data and demographics data. The multimodal customer features are composed of dense numerical features (such as number of orders, GMV, number of visits) and sparse categorical features (such as gender, education, occupation, location). For low dimensional categorical features, we can use one-hot encoding to transform them into numbers, for example gender and education level. The location feature contains street address, city name, state name and zip code, so it's a very high dimensional categorical feature, which is too inefficient to use one-hot encoding. We convert the location feature into location embedding using transfer learning and then concatenate it with other features together, which is explained in the next part.

4.3 Embedding Model Structure

The embedding model is described in Figure 3. We have the transaction, visit, engagement, demographics and location features as input to the embedding model. The location features are treated as textual sentences and then transformed to location embeddings using transfer learning of pre-trained text embedding models. Then we concatenate the numerical features and location embeddings as input to the final feed-forward neural network. The feed-forward network is composed of multiple fully connected ReLU layers. The output of the feed-forward network is a 128 dimensional embedding as customer representation.

4.3.1 Location Embedding. We tried a few different approaches to convert location text to location embedding. One approach is to use a pre-trained word embedding model in PyTorch (GloVe), which is illustrated in Figure 4. Another approach is to use the state-of-the-art BERT model [2] for text representation learning, which is shown in Figure 5. We also fine tune the pre-trained BERT model in our training process.
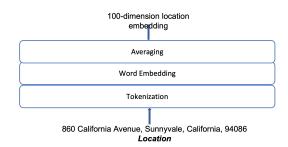
100-dimension location
embedding

Averaging

Word Embedding

Tokenization

860 California Avenue, Sunnyvale, California, 94086
Location

Figure 4: Location Embedding using Word Embedding.

5 EXPERIMENTAL RESULTS

For evaluation, we setup the training, validation and testing datasets by different time windows. For example, we can use data of last *n* years as training data, data of next one month or one quarter as validation and testing data. The evaluation metric for model quality is mean absolute error (MAE). Due to Walmart's privacy policy,
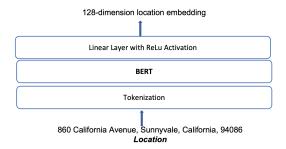
128-dimension location embedding

| Linear Layer with ReLu Activation |
| BERT |
| Tokenization |

860 California Avenue, Sunnyvale, California, 94086
Location

Figure 5: Location Embedding using BERT.

the results are presented as percentage proportions to the baseline embedding model.

In this section, we compare the quality and inference time of different model setups, from different numbers of fully connected layers to different location embedding methods. The baseline model is using two fully connected ReLU layers and no location embedding. The results are shown in Table 1 and Table 2.

Table 1: Quality (MAE) of Embedding Model with Different Setups.

	No Location	Word Embedding	BERT
2 layers	100%	97%	91%
3 layers	94%	87%	86%

Table 2: Inference Time of Embedding Model with Different Setups.

	No Location	Word Embedding	BERT
2 layers	100%	110%	>500%
3 layers	105%	115%	>500%

In both Table 1 and Table 2, lower percentage indicates better model performance. Although embedding model with BERT is the best one in terms of model quality, its inference latency increases by 4 times compared to baseline, which is very slow. Considering we need to process hundreds of millions of customers offline, the total running time of our system using BERT is not ideal. Embedding model with word embedding strikes a good balance between quality and inference latency.

6 CONCLUSIONS

In this paper, we propose an effective and efficient marketing audience expansion system, which can handle hundreds of millions of customers. We use a deep learning model to generate customer embeddings. The deep learning model can handle both dense numerical features and sparse categorical features. And the model encodes location embedding using transfer learning Our system has great adaptability by constructing different similarity metrics for different campaigns. The similarity metrics are interpretable and meaningful and can reflect different business metrics. Extensive

experiments have been conducted to demonstrate the scalability, efficiency and quality of our system and embedding model.

There are several directions for the future work in our marketing audience expansion system. First, we can explore the direction of combining both similarity-based and classification-based approaches for searching lookalike customers. Second, we can study how to filter and rank the lookalike customers using machine learning models to improve ranking quality.

REFERENCES

[1] S Amirrahmat, KA Alshibli, MF Jarrar, B Zhang, and RA Regueiro. 2018. Equivalent continuum strain calculations based on 3D particle kinematic measurements of sand. *International Journal for Numerical and Analytical Methods in Geomechanics* 42 (2018), 999–1015. Issue 8. https://doi.org/10.1002/nag.2779

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805

[3] Stephanie deWet and Jiafan Ou. 2019. Finding users who act alike: transfer learning for expanding advertiser audiences. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2251–2259.

[4] Dihong Gong, Daisy Zhe Wang, and Yang Peng. 2017. Multimodal Learning for Web Information Extraction. In *Proceedings of the 25th ACM International Conference on Multimedia (MM '17)*. Association for Computing Machinery, New York, NY, USA, 288–296.

[5] Mihajlo Grbovic and Haibin Cheng. 2018. Real-time personalization using embeddings for search ranking at airbnb. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 311–320.

[6] Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating Large-Scale Inference with Anisotropic Vector Quantization. In *International Conference on Machine Learning*. https://arxiv.org/abs/1908.10396

[7] Robert A. Jarrow, Rinald Murataj, Martin T. Wells, and Liao Zhu. 2021. The Low-volatility Anomaly and the Adaptive Multi-Factor Model. *arXiv preprint arXiv:2003.08302* (2021).

[8] Jinling Jiang, Xiaoming Lin, Junjie Yao, and Hua Lu. 2019. Comprehensive audience expansion based on end-to-end neural prediction. In *CEUR Workshop Proceedings*, Vol. 2410. CEUR Workshop Proceedings.

[9] Cheng Jie, Zigeng Wang, Da Xu, and Wei Shen. 2022. Multi-objective Cluster Based Bidding Algorithm for E-Commerce Search Engine Marketing System. *Frontiers in Big Data* (2022), 77.

[10] Cheng Jie, Da Xu, Zigeng Wang, and Wei Shen. 2022. Deep Learning Based Page Creation for Improving E-Commerce Organic Search Traffic. https://doi.org/10.48550/ARXIV.2209.10792

[11] Cheng Jie, Da Xu, Zigeng Wang, Lu Wang, and Wei Shen. 2021. An Efficient Group-based Search Engine Marketing System for E-Commerce. https://doi.org/10.48550/ARXIV.2106.12700

[12] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[13] Yifei Li, Kuangyan Song, Yiming Sun, and Liao Zhu. 2021. FrequentNet: A Novel Interpretable Deep Learning Model for Image Classification. *arXiv preprint arXiv:2001.01034* (2021).

[14] Haishan Liu, David Pardoe, Kun Liu, Manoj Thakur, Frank Cao, and Chongzhe Li. 2016. Audience expansion for online social network advertising. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 165–174.

[15] Huiyang Luo, Huiluo Chen, Runyu Zhang, Yao Ren, Boning Zhang, Richard A. Regueiro, Khalid Alshibli, and Hongbing Lu. 2022. 4 - Constitutive behavior of granular materials under high rate of uniaxial strain loading. In *Advances in Experimental Impact Mechanics*, Bo Song (Ed.). Elsevier, 99–124. https://doi.org/10.1016/B978-0-12-823325-2.00005-4

[16] Qiang Ma, Musen Wen, Zhen Xia, and Datong Chen. 2016. A Sub-linear, Massive-scale Look-alike Audience Extension System A Massive-scale Look-alike Audience Extension. In *Proceedings of the 5th International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications at KDD 2016 (Proceedings of Machine Learning Research)*, Wei Fan, Albert Bifet, Jesse Read, Qiang Yang, and Philip S. Yu (Eds.), Vol. 53. PMLR, San Francisco, California, USA, 51–67. https://proceedings.mlr.press/v53/ma16.html

[17] Morteza Shahriari Nia, Christan Grant, Yang Peng, Daisy Zhe Wang, and Milenko Petrovic. 2013. University of Florida Knowledge Base Acceleration Notebook. *The Twenty-Second Text REtrieval Conference (TREC 2013)* (2013).

[18] Morteza Shahriari Nia, Christan Earl Grant, Yang Peng, Daisy Zhe Wang, and Milenko Petrovic. 2014. Streaming Fact Extraction for Wikipedia Entities at

Web-Scale.. In *FLAIRS Conference*.

[19] Aditya Pal, Chantat Eksombatchai, Yitong Zhou, Bo Zhao, Charles Rosenberg, and Jure Leskovec. 2020. Pinnersage: Multi-modal user embedding framework for recommendations at pinterest. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2311–2320.

[20] Yang Peng. 2017. Multimodal Fusion: A Theory and Applications. *University of Florida* (2017).

[21] Yang Peng. 2023. Query-Driven Knowledge Graph Construction Using Question Answering and Multimodal Fusion. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 1119–1126. https://doi.org/10.1145/3543873.3587567

[22] Yang Peng and Daisy Zhe Wang. 2022. Knowledge Base Completion using Web-Based Question Answering and Multimodal Fusion. https://doi.org/10.48550/ARXIV.2211.07098

[23] Yang Peng and Daisy Zhe Wang. 2022. Query-Driven Knowledge Base Completion using Multimodal Path Fusion over Multimodal Knowledge Graph. https://doi.org/10.48550/ARXIV.2212.01923

[24] Yang Peng, Daisy Zhe Wang, Ishan Patwa, Dihong Gong, and Chunsheng Victor Fang. 2015. Probabilistic Ensemble Fusion for Multimodal Word Sense Disambiguation. In *Multimedia (ISM), 2015 IEEE International Symposium on*. IEEE, 172–177.

[25] Yang Peng, Xiaofeng Zhou, Daisy Zhe Wang, and Chunsheng Victor Fang. 2016. Scalable image retrieval with multimodal fusion. In *The Twenty-Ninth International Flairs Conference*.

[26] Yang Peng, Xiaofeng Zhou, Daisy Zhe Wang, Ishan Patwa, Dihong Gong, and Chunsheng Fang. 2016. Multimodal Ensemble Fusion for Disambiguation and Retrieval. *IEEE MultiMedia* (2016).

[27] Christopher T Senseney, Zheng Duan, Boning Zhang, and Richard A Regueiro. 2017. Combined spheropolyhedral discrete element (DE)-finite element (FE) computational modeling of vertical plate loading on cohesionless soil. *Acta Geotechnica* 12 (2017), 593–603. https://doi.org/10.1007/s11440-016-0519-8

[28] Boning Zhang, Eric B. Herbold, Michael A. Homel, and Richard A. Regueiro. 2015. *DEM Particle Fracture Model*. Technical Report. Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States).

[29] Boning Zhang, Richard A. Regueiro, Andrew M. Druckrey, and Khalid Alshibli. 2018. Construction of poly-ellipsoidal grain shapes from SMT imaging on sand, and the development of a new DEM contact detection algorithm. *Engineering Computations* 35, 2 (2018), 733–771. https://doi.org/10.1108/EC-01-2017-0026

[30] Liao Zhu. 2020. *The Adaptive Multi-Factor Model and the Financial Market*. eCommons.

[31] Liao Zhu, Sumanta Basu, Robert A. Jarrow, and Martin T. Wells. 2020. High-Dimensional Estimation, Basis Assets, and the Adaptive Multi-Factor Model. *The Quarterly Journal of Finance* 10, 04 (2020), 2050017.

[32] Liao Zhu, Robert A. Jarrow, and Martin T. Wells. 2021. Time-Invariance Coefficients Tests with the Adaptive Multi-Factor Model. *The Quarterly Journal of Finance* 11, 04 (2021), 2150019.

[33] Liao Zhu, Ningning Sun, and Martin T. Wells. 2021. Clustering Structure of Microstructure Measures. *arXiv preprint arXiv:2107.02283* (2021).

[34] Liao Zhu, Haoxuan Wu, and Martin T. Wells. 2021. A News-based Machine Learning Model for Adaptive Asset Pricing. *arXiv preprint arXiv:2106.07103* (2021).