A Dataset and Baselines for e-Commerce Product Categorization

Yiu-Chang Lin, Pradipto Das
Rakuten Institute of Technology – Americas, Boston, USA
{yiuchang.lin,pradipto.das}@rakuten.com

Andrew Trotman
University of Otago
andrew@cs.otago.ac.nz

Surya Kallumadi
The Home Depot
surya@ksu.edu

ABSTRACT

We make available a document collection of a million product titles from 3,008 anonymized categories of the `rakuten.com` product catalog. The anonymization has been done due to intellectual property rights on the underlying *data organization* taxonomy. Our analysis of the characteristics of the 800,000 training and 20,000 validation titles show that they match the test set of 180,000 titles. Twenty six independent teams participated in an automatic product categorization challenge on this dataset. We present results and analysis and suggest strong baselines for this collection and task.

CCS CONCEPTS

• **Information systems → Clustering and classification**;

KEYWORDS

Document Collection, e-Commerce, Taxonomy Categorization

ACM Reference Format:
Yiu-Chang Lin, Pradipto Das, Andrew Trotman, and Surya Kallumadi. 2019. A Dataset and Baselines for e-Commerce Product Categorization. In *The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19), October 2–5, 2019, Santa Clara, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3341981.3344237

1 INTRODUCTION

Taxonomy categorization of product listings is a fundamental problem for any e-commerce platform and has applications ranging from basic data organization and personalized search recommendation, to query understanding and targeted campaigning. Manual and rule based approaches to categorization are error prone and expensive [15] because commercial product taxonomies have thousands of leaf nodes with semantically similar paths to the root. Academic advances on this task have been limited by a lack of real-world data from a commercial e-commerce platform.

We are making available a real-world data set with one million product listings from 3,008 categories. Twenty six teams (from academic and from industrial backgrounds) participated in the automatic product categorization data challenge that was run on this data. This paper presents our analysis of the data we provide, and strong baselines for further research on this task.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICTIR '19, October 2–5, 2019, Santa Clara, CA, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6881-0/19/10…$15.00
https://doi.org/10.1145/3341981.3344237

Product Titles	Category path
Replacement Viewsonic VG710 LCD Monitor 48Watt AC Adapter 12V 4A	3292>114>1231
Ka-Bar Desert MULE Serrated Folding Knife	4238>321>753>3121
5.11 TACTICAL 74280 Taclite TDU Pants, R/M, Dark Navy	4015>3285>1443>20
Skechers 4lb S Grip Jogging Weight set of 2- Black	2075>945>2183>3863
Generations Small Side Table White	4015>3636>1319>1409>3606

Table 1: Examples from the training set. Category paths are anonymized to retain intellectual property.

Product Titles
Disc Brake Caliper Guide Pin Boot Kit Front Carlson 16137
Wire Shelf, Green ,Metro, 2442NK3
Parallel Lines Velvet Cushion
GROZ 36JN79 Filter Element, 40 Microns, Intermediate
Chenille Kraft Wonderfoam Magnetic Alphabet Letters, Assorted Colors. 105/Pack - CKC4357

Table 2: Examples from the test set.

2 RELATED WORK

There have been many discussions about the immense importance of ontologies and taxonomies for e-commerce [5], and we concur with their conclusions. The problem of assigning products to taxonomy has been addressed for some time [2, 4] and anecdotally, this is a problem currently faced by large scale e-commerce platforms.

We are not the first to release data for product classification, or to run a challenge on such data. In 2015, The Otto Group (which includes Crate & Barrel) released a training set of 61,879 products and a test set of 144,369 products on Kaggle[1]. That dataset has product listings represented as a set of ninety three strictly numeric features and the task is to categorize them into nine categories that represented top level categories in their taxonomy tree. Evaluation was with multi-class logarithmic loss. Our data significantly differs from theirs, for instance, we include one million product titles organized into an *organizational taxonomy* of 3,008 leaf nodes.

Data challenges in a related area have also occurred, for instance, Schulten et al. [12] present the challenge of taxonomy mapping. This is an important and active area of research [1, 7], that is different from the one presented here.

McAuley et al. [11] provide a crawl of Amazon's product pages including 142.8 million reviews[2], but do not run a data challenge. They address recommendation of substitute products (e.g. would you prefer this phone to that phone) and complementary items (e.g. do you need batteries with that). A *navigational taxonomy* could be extracted from that data, but we are interested in an *organizational taxonomy*, which is normally proprietary. Our dataset contains the organizational taxonomy labels for each listing, albeit anonymized.

3 DATA SET

A large-scale e-commerce platform usually handles millions of product listings on a daily basis. Data of this scale is difficult to

[1] https://www.kaggle.com/c/otto-group-product-classification-challenge
[2] http://jmcauley.ucsd.edu/data/amazon/

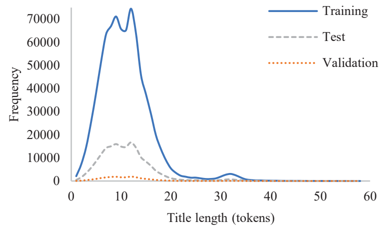

Figure 1: Title length distributions (in tokens).

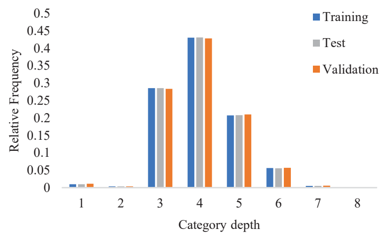

Figure 2: Lengths of category paths (proportion of data set).

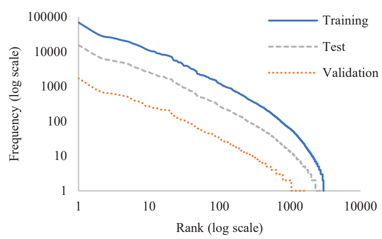

Figure 3: Frequencies of each of the 3008 category (ordered from most to least frequent).

work with, and so in January 2018, we sorted and deduplicated the product titles of the full catalog snapshot from `rakuten.com`[3] and then randomly sampled without replacement, exactly one million product listings.

We partition this dataset into a training set of 800, 000, a validation set of 20, 000, and a test set of 180, 000 listings using category-wise stratified sampling. Each listing consists of a product title and taxonomy tree path to the leaf node (the category). To preserve intellectual property, labels in the taxonomic tree are replaced with random integers and only those parts of the tree that are covered by the million listings are released. The anonymization does not change the nature of the problem of taxonomy categorization.

Each listing, therefore, consists of a textual name and a sequence of integers representing the path (left to right, less to more specific) through the taxonomy. The two fields are tab separated. Listings only occur at leaves, never at internal nodes. The test and validation sets contain only titles (the object of the data challenge was to predict the path). The gold standard contains both the product titles and the categories. Table 1 and Table 2 show examples of product titles from the training and test set, respectively.

3.1 Data Characteristics

Table 3 presents the basic statistics of our dataset. Once the titles are tokenized on white space, the training set of 800, 000 titles ranges in length from 1 to 58 tokens with an average length of 11 tokens. Figure 1 illustrates the uneven distribution of title lengths in tokens

[3] The dataset can be downloaded from https://forms.gle/acBxFo3Qwdi1Edyy5

Parameter	Training	Validation	Test
Titles (Documents)	800, 000	20, 000	180, 000
Length in bytes	1 − 400	2 − 256	1 − 258
Avg length in bytes	68.75	68.83	68.80
Length in tokens	1 − 58	1 − 49	1 − 58
Avg length in tokens	10.93	10.93	10.94
Path length	1 − 8	1 − 7	1 − 8

Table 3: Statistics of each part of the collection.

in the three sets. From visual inspection, the three lines are similar in shape – and indeed both the validation set and the training set show a Pearson's correlation of > 0.99 with the test set. Due to the use of stratified sampling, the distributions over the three sets are similar. The titles are unevenly distributed across 3, 008 categories with a minimum depth of 1, a maximum depth of 8, and an average depth of 4. Figure 2 presents the proportion of titles according to their path lengths in the taxonomy. Again due to stratified sampling, the three sets are similar.

In each of the three sets the top ten categories compose $\approx 30\%$ of the titles, and the top 37 categories compose $\approx 50\%$ of the titles. Figure 3 shows the distribution of titles across all the leaf-level categories (ordered by frequency). From visual inspection, the three lines are similar in shape – and indeed both the validation set and the training set show a Pearson's correlation of > 0.99 with the test set.

This section has presented the statistics of the three sets in the collection and shown that the training set and the validation set both very strongly correlate with the test set. We believe that this dataset, for taxonomy categorization on real world product listings, leads itself to strong baselines being developed, and is suited to furthering the research in this area.

4 EVALUATION

As a method of obtaining a reasonable baseline for automatic product categorization, and as a way to assess the difficulty of this task, we ran a data challenge at the ACM SIGIR 2018 Workshop on e-commerce. This section describes our experiments.

4.1 Timeline for System Submissions

We started promoting the challenge on 24th March, 2018 and immediately saw several registrations (10 in the first 24 hours).

Participants were permitted to submit up to three runs per day between 9th April, 2018 and 23rd June, 2018. These runs were evaluated on the validation set. On 24th June, 2018 submissions were finalized and the runs were subsequently evaluated on the test set. Results are presented in Section 5.

4.2 Metrics

The distribution of listings over the taxonomy is highly skewed (see Figure 3). In an industrial e-commerce setting, usually, classification performance on the long tail is not considered to be important. For this reason, we choose the weighted versions of precision, recall, and $F1$ as indicators of system performance. A predicted taxonomy path is considered correct if and only if it *exactly matches* the taxonomy path in the gold standard (partial matches are considered to be incorrect). Denote by K, a total number of classes, $\{c_i | i = 1, 2, ..., K\}$

in the test set. The number of true instances for each class (support) is n_i, and the total number of instances is $N = \sum_{i=1}^{K} n_i$. If we compute the precision (P_i), recall (R_i) and F1 ($F1_i$) for each class c_i, then the weighted metrics are:

$$P_w = \sum_{i=1}^{K} \frac{n_i}{N} P_i \qquad R_w = \sum_{i=1}^{K} \frac{n_i}{N} R_i \qquad F1_w = \sum_{i=1}^{K} \frac{n_i}{N} F1_i$$

We only report the low Macro-F1 ($= \frac{1}{K} \sum_{i=1}^{K} F1_i$) numbers in tables 4 and 5 to highlight the long tail problem.

5 RESULTS

Twenty six teams participated in the challenge (we do not list them due to space limits). Each team submitted their predicted categories on the validation set and the test sets in the same tab-separated format used with the training set (see Table 1). During the challenge, runs were scored against the validation set using weighted precision, recall and F_1 as defined in Section 4.2. A leader board on the data challenge website[4] tracked the submission scores.

At the end of the challenge, the *final* run for each team was evaluated on the test set. Although the performance on the validation set could be known during the challenge, the performance on the test set could not be known until the challenge had finished. Table 4 shows the leader board at the end of the challenge, giving the final scores on the validation set, while Table 5 shows the final scores on the test set computed after the challenge.

The top five teams on the validation set as shown on the leader board are **mcskinner**, **MKANEMAS**, **tiger**, **Uplab** and **JCWRY** with 0.8510, 0.8421, 0.8404, 0.8375 and 0.8278 $F1_w$ scores respectively. The same five teams rank top and in the same order on the test set with $F1_w$ scores of 0.8510, 0.8397, 0.8379, 0.8364 and 0.8295 respectively. Although there is no change in the rank of the top five teams between validation and test, changes are seen lower down in the ranks (amongst others, for instance, **Uplab-2** changes rank). We computed a Pearson's correlation coefficient of > 0.99 on the absolute $F1_w$ scores and a Spearman's rank correlation coefficient of > 0.98 on the rank ordering of systems. This very high correlation suggests that the validation set is an excellent indicator of expected performance on this task.

6 SYSTEM DESCRIPTIONS

In this section we outline the approaches taken by teams that performed best on the validation and test sets, ordered by the performance on the test set.

- **mcskinner** ($F1_w$: 0.8510) achieved the highest scores. The system uses an ensemble of LSTMs and show a positive impact of dense connections between recurrent and output layers through the use of pooling layers. Their final solution is produced using a bidirectional ensemble of six LSTMs with a balanced pooling view architecture [13].
- **MKANEMAS** ($F1_w$: 0.8397) formulate the task as a simple classification problem of just leaf categories. The key feature of their system is the combination of a convolutional neural network and bidirectional LSTM using ad-hoc features generated from an external data set [14].

[4]We also made available a script that computes these scores. https://github.com/sigir-ecom/dataChallenge

Team	P_w	R_w	$F1_w$	Macro-$F1$
mcskinner	0.8734	0.8425	0.8510	0.4999
MKANEMAS	0.8509	0.8445	0.8421	0.4994
tiger	0.8552	0.8389	0.8404	0.4881
Uplab	0.8435	0.8427	0.8375	0.4902
JCWRY	0.8545	0.8172	0.8278	0.4670
neko	0.8311	0.8296	0.8245	0.4717
Ravenclaw	0.8394	0.8118	0.8197	0.3939
ssdragon	0.8310	0.8173	0.8185	0.4068
RITB-Baseline	0.8389	0.8097	0.8172	0.3909
inception	0.8364	0.8087	0.8166	0.3860
Uplab-2	0.8196	0.8228	0.8149	0.4621
minimono	0.8119	0.8042	0.8020	0.3782
Tyche	0.8536	0.7655	0.7976	0.0538
Topsig	0.8009	0.8042	0.7967	0.4240
VanGuard	0.7950	0.7915	0.7871	0.3233
Waterloo	0.7819	0.7853	0.7767	0.4065
CorUmBc	0.7822	0.7722	0.7702	0.3728
Sam-chan	0.7695	0.7704	0.7617	0.3636
Tyken2018	0.7545	0.7561	0.7431	0.3415
Or	0.7446	0.7232	0.7226	0.2925

Table 4: Final results on the validation set of 20, 000 titles ordered by $F1_w$ scores for the top twenty systems.

Team	P_w	R_w	$F1_w$	Macro-$F1$
mcskinner	0.8693	0.8417	0.8510	0.4989
MKANEMAS	0.8423	0.8425	0.8397	0.4992
tiger	0.8398	0.8429	0.8379	0.4893
Uplab	0.8367	0.8418	0.8364	0.4881
JCWRY	0.8531	0.8172	0.8295	0.4696
neko	0.8268	0.8306	0.8256	0.4732
Ravenclaw	0.8291	0.8114	0.8174	0.3922
Uplab-2	0.8188	0.8245	0.8174	0.4629
ssdragon	0.8229	0.8162	0.8172	0.4061
RITB-Baseline	0.8276	0.8075	0.8140	0.3894
inception	0.8261	0.8076	0.8138	0.3852
Tyche	0.8597	0.7643	0.8001	0.0572
minimono	0.8016	0.8021	0.7991	0.3804
Topsig	0.7919	0.8011	0.7937	0.4235
VanGuard	0.7902	0.7917	0.7885	0.3282
HSJX-ITEC-YU	0.7807	0.7819	0.7787	0.4192
Waterloo	0.7803	0.7858	0.7780	0.4076
CorUmBc	0.7744	0.7711	0.7689	0.3726
Sam-chan	0.7721	0.7749	0.7669	0.3654
Tyken2018	0.7658	0.7608	0.7514	0.3444

Table 5: Final results on the test set of 180, 000 titles ordered by $F1_w$ scores for the top twenty systems.

- **tiger** ($F1_w$: 0.8379) combine multiple models based on single-label and multi-level label predictions, as well as characteristics of the taxonomy tree structure. The training data set and the validation data set are merged to pre-train word vectors for calculating semantic similarity. To address high category imbalance, sampling and data enhancement techniques are used. They build eight sample data sets according to the category hierarchy and develop two classification algorithms to build models for different levels and search paths using category trees [16].

- **Uplab** submitted three systems based on different classifier types, including single flat linear support vector machines classifier ($F1_w$: 0.8364), a top down ensemble which combines top-level and sub-level classifiers ($F1_w$:0.8174) and a CNN with pre-trained word embeddings ($F1_w$: 0.6511). They found that *tf-idf* weighting with both bi-gram and unigram features work best for categorization [8].
- **JCWRY** ($F1_w$: 0.8295) use deep convolutional neural networks with oversampling, threshold moving and error correct output coding to predict product taxonomies. Their highest accuracy was obtained through an ensemble of multiple networks, such as Kim-CNN and Zhang-CNN, trained on different extracted features inputs, including doc2vec, Named Entity Recognition and Parts of Speech features [9].

6.1 Confidence Intervals

We use bootstrap, a re-sampling strategy [6, 10], to estimate the confidence intervals for the weighted F1 scores of the top twenty systems. The basic principle of the bootstrap is to evaluate the properties of an arbitrary estimator $\theta(y_1, ..., y_n)$, through the empirical cumulative distribution function (cdf) of the sample $Y_1, ..., Y_n$, $F_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{Y_i \leq y}$, instead of the theoretical cdf F. More precisely, $\theta(F_n) = \int h(\mathrm{y})dF_n(\mathrm{y})$ is an obvious estimator to estimate $\theta(F) = \int h(\mathrm{y})dF(\mathrm{y})$ for any continuous function h.

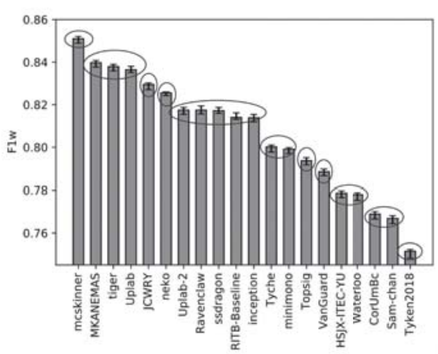

Figure 4: 95% confidence intervals for top twenty teams.

For each submitted system, θ is the estimator for the weighted F1 function, $h(\mathrm{y})$, computed over the set of n = 180,000 test titles, where y_is are the binary indicators for correct predictions for the test titles x_is. When the X_is and hence the Y_is are independent and identically distributed random variables, as in our case, the Glivenko-Cantelli theorem [3] states that $F_n(\mathrm{y})$ converges in probability to F and hence $\theta(F_n)$ is a consistent estimator for $\theta(F)$.

The bootstrap estimator becomes $\theta_B(F_n(\mathrm{y})) = \frac{1}{B}\sum_{b=1}^{B} h(\mathrm{Y}_b^*)$, where Y_b^* is a sampling with replacement of Y and B is the number of bootstrap samples. We set $B = 1,000$. The bias of the estimator, $E_{F_n}[\theta_B(F_n(\mathrm{y})) - \theta(F_n(\mathrm{y}))]$ is used to calculate the confidence interval of the estimator for $\theta(F)$. The confidence interval $[\theta(F_n) - \alpha, \theta(F_n) - \beta]$ on $\theta(F_n)$ is constructed by imposing the constraint $p_{F_n}(\alpha \leq \theta_B(F_n) - \theta(F_n) \leq \beta) = c$ on (α, β), where c is the desired confidence level, which is our case is 0.95.

For each system, we re-sample the predictions (with replacement) for a total of $B = 1,000$ times. For each of the b re-samplings, we keep track of the weighted F1 scores and sort the B biases in ascending order. Therefore, the lower and upper bounds of the confidence interval are determined by subtracting from $\theta(F_n)(= F1_w)$, the $\alpha = 2.5^{th}$ percentile and $-\beta = 97.5^{th}$ percentile values from the sorted array. Figure 4 shows the confidence intervals and clustering of the top twenty systems. The clusters are shown with black ovals. Within a cluster, differences between systems are statistically insignificant based on a confidence level of 95%.

7 CONCLUSIONS

With this paper we have released a set of a million product titles from 3,008 categories of `rakuten.com`. Our analysis shows that the characteristics of the training set and the validation set closely match those of the test set. A successful data challenge saw twenty six teams from academia and industry compete. The highest performing team achieving a weighted $F1$ score of 0.8510 on the test set – which we consider to be a high-mark baseline for automatic product categorization on this collection.

This data presents several *additional* research avenues. Such tasks include designing better classifiers that address the long tail problem, topic modeling over a taxonomy, or even a minimally supervised attribute extraction from product titles.

REFERENCES

[1] SS Aanen, D Vandic, and F Frasincar. 2015. Automated Product Taxonomy Mapping in an e-Commerce Environment. *Expert Syst. Appl.* 42, 3 (2015), 1298–1313.
[2] S Abels and A Hahn. 2005. Automatic Classification and Re-classification of Product Data in e-Business. In *The 2005 Symposium on Applications and the Internet.* 350–353.
[3] P Billingsley. 1986. *Probability and Measure* (second ed.). John Wiley and Sons.
[4] E Cortez, M Rojas Herrera, AS da Silva, ES de Moura, and M Neubert. 2011. Lightweight Methods for Large-Scale Product Categorization. *JASIS&T* 62, 9 (2011), 1839–1848.
[5] Y Ding, D Fensel, M Klein, B Omelayenko, and E Schulten. 2004. The Role of Ontologies in eCommerce. In *Handbook on ontologies.* 593–615.
[6] B Efron and R Tibshirani. 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statist. Sci.* 1, 1 (02 1986), 54–75.
[7] VR Embar, G Farnadi, J Pujara, and L Getoor. 2018. Aligning Product Categories using Anchor Products. In *First Workshop on Knowledge Base Construction, Reasoning and Mining.*
[8] S Goumy and M-A Mejri. 2018. Ecommerce Product Title Classification. In *SIGIR 2018 Workshop on eCommerce.*
[9] Y Jia, X Wang, H Cao, B Ru, and T Yang. 2018. An Empirical Study of Using An Ensemble Model in E-commerce Taxonomy Classification Challenge. In *SIGIR 2018 Workshop on eCommerce.*
[10] EL Lehmann. 1963. Nonparametric Confidence Intervals for a Shift Parameter. *Ann. Math. Statist.* 34, 4 (12 1963), 1507–1512.
[11] J McAuley, R Pandey, and J Leskovec. 2015. Inferring Networks of Substitutable and Complementary Products. In *KDD 2015.* 785–794.
[12] E Schulten, H Akkermans, G Botquin, M Dörr, N Guarino, N Lopes, and N Sadeh. 2001. Call for Participants: The e-Commerce Product Classification Challenge. *IEEE Intelligent systems* 4 (2001), 86–c3.
[13] M Skinner. 2018. Product Categorization with LSTMs and Balanced Pooling Views. In *SIGIR 2018 Workshop on eCommerce.*
[14] S Suzuki, Y Iseki, H Shiino, H Zhang, A Iwamoto, and F Takahashi. 2018. Convolutional Neural Network and Bidirectional LSTM Based Taxonomy Classification Using External Dataset at SIGIR eCom Data Challenge. In *SIGIR 2018 Workshop on eCommerce.*
[15] B Wolin. 2002. Automatic Classification in Product Catalogs. In *SIGIR '02.* 351–352.
[16] W Yu, Z Sun, H Liu, Z Li, and Z Zheng. 2018. Multi-level Deep Learning based E-commerce Product Categorization. In *SIGIR 2018 Workshop on eCommerce.*