Enhanced Text Classification using Proxy Labels and Knowledge Distillation

Rohan Sukumaran
IIIT Sri City
Sri City, Chittoor, India
rohan.s16@iiits.in

Sumanth Prabhu
Applied Research, Swiggy
Bangalore, India
sumanth.prabhu@swiggy.in

Hemant Misra
Applied Research, Swiggy
Bangalore, India
hemant.misra@swiggy.in

ABSTRACT

Text Classification has a variety of applications in the pickup and delivery services industry where customers require one or more items to be picked up from a location and delivered to a certain destination. Categorizing these customer transactions helps understand the market needs and trends while also assisting in building a personalized experience for each customer segment. In this paper, each transaction is accompanied by a free text description provided by the customer to describe the products to be picked up and delivered. These descriptions tend to be short, incoherent and code-mixed (Hindi-English) text.Here, we focus on a specific use-case where each customer transaction can be mapped to a single product category. We propose a cost-effective transaction classification approach based on proxy-labelling and knowledge distillation using the transaction descriptions provided by the customer. We introduce R-ALBERT, a model trained with RoBERTa as the "teacher" and ALBERT (33x fewer parameters than RoBERTa) as the "student". Further, we benchmark R-ALBERT on a large internal dataset as well as the 20Newsgroup dataset. We see that our model shows a 2% increase in performance with 33x fewer parameters. The model is currently deployed in production and is helping understand the customer behaviour across product categories and customer segments.

KEYWORDS

Proxy Labelling, Text Classification, Knowledge Distillation, Semi-supervised Learning

ACM Reference Format:

Rohan Sukumaran, Sumanth Prabhu, and Hemant Misra. 2022. Enhanced Text Classification using Proxy Labels and Knowledge Distillation. In *5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD) (CODS-COMAD 2022), January 8–10, 2022, Bangalore, India.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3493700.3493722

1 INTRODUCTION

Text classification is a classical problem in Natural Language Understanding (NLU) with applications in Question Answering [15], Sentiment Analysis [38], Intent Classification [21] and many other similar tasks. The advancement in machine learning has enhanced the scale of adaption of such capabilities to enable richer customer experience. However, applying these capabilities in an industry setting, particularly in a supervised setting, can prove to be challenging. More concretely, typical supervised settings demand the availability of abundant high quality labelled training data. Manually labelling this data can be expensive. Thus, weaker forms of supervision [26, 27, 31] was explored to label data in a cost effective manner. In this regard, semi-supervision [20] has proven to be particularly useful in generating *proxy labels* for unlabeled data, given the availability of a relatively smaller set of labelled data.

Deep Learning models achieve state-of-the-art results on GLUE [40], RACE [17] and SQuAD [29, 30] benchmarks. With the advent of Transfer Learning [37], large deep learning based models perform well even with access to minimal training data. However, memory and inference time constraints make deploying such models challenging in a real-time, resource constrained industry setting [7]. As a result, various techniques have been explored to perform model compression [7] with minimal loss of information.

In this paper, we consider a text classification use-case specific to pickup and delivery services, where customers make use of short phrases to describe the products to be picked up from a certain location and dropped at a target location. Table 1 shows a few examples of the descriptions used by our customers to describe their transactions. Customers tend to use short code-mixed (using more than one language in the same message[1]) and incoherent textual descriptions of the products for describing them in the transaction. These descriptions, if mapped to a fixed set of categories, will help assist critical business decisions such as - enhancing the customer experience on the platform, understanding the importance of each category and issues faced in them, demography driven prioritization of categories, launch of new product categories and more. Furthermore, a transaction may comprise of multiple products which adds to the complexity of the task. In this work, we focus on a multi-class classification of transactions, where a single majority category drives the transaction.

Due to the incoherent and code-mixed nature of the transaction descriptions, we explored *supervised classification* of transaction types and this required labelled training data. The training data used in this paper was labeled manually by subject matter experts (SMEs). This was proving to be a very expensive exercise, and hence, necessitated exploration of weak labelling strategies to ensure cost effective development of models. Our experiments with multiple Deep Learning models revealed that RoBERTa ([22]) was the best performing model for our task. However, owing to the large number of parameters, it was not feasible to deploy this model at production

CODS-COMAD 2022, January 8–10, 2022, Bangalore, India

ACM ISBN 978-1-4503-8582-4/22/01…$15.00
https://doi.org/10.1145/3493700.3493722

[1] In our case, Hindi written in roman script is mixed with English

Transaction Description	Category
"Get me my lehanga" **Translation :** Get me my skirt	Clothes
"Buy a 500gm packet of Whole grain Atta"	Grocery
"Get a roll of paratha"	Food
"Mera do bags leaoo" **Translation :** Bring two of my bags	Package

Table 1: Samples of actual transaction descriptions used by our customers along with their corresponding categories as labelled by the subject matter experts (SMEs).

scale. Furthermore, we observed that the lighter versions of BERT such as ALBERT (base) [18, 41] though production friendly did not match the performance of RoBERTa.

To address the challenges described above, we showcase: a) an approach that leverages proxy labelling via semi-supervision to reduce the manual labeling cost and, b) also explore knowledge distillation to build a smaller model (in terms of numbers of parameters) that matches the performance of the SOTA heavier models such as RoBERTa. The key contributions of this paper are:

- **Weak Labelling:** A proxy labelling framework based on semi-supervised learning to reduce the cost of labelling data.
- **Knowledge Distillation Framework:** Training a lightweight model (33x lesser parameters) with the help of weak labels, which is able to match the performance of a much heavier model.

2 RELATED WORKS

Text classification problems with code-mixed inputs have been studied and transformer based models perform well on benchmarks [4, 23] like TREC-6 [39] and DBpedia [2].

Weak labeling of data Text classification based on proxy labelling has become a popular practice to achieve low cost model training. [11] explored approaches based on topic modelling to predict labels for documents. Our problem setting involves short transaction descriptions that do not perform well with standard topic modelling techniques. [19] worked with unlabeled data by identifying a minimal set of seed word based pseudo labels for documents and trained a Naive Bayes model using semi-supervision. Our problem setting is focused on leveraging manually tagged data as well as unlabeled data to improve the performance of the model. [42] proposed a semi-supervised pipeline leveraging unlabeled data to improve performance of state-of-the-art (SOTA) models in image classification. We extend this idea to an NLP problem setting focused on improving the performance of BERT-based models. [25, 43] explore self training where a model's own predictions on unlabelled data are leveraged to expand the training data. We apply a similar approach with the difference that our light weight model learns from state-of-the-art heavier models. There are many more promising approaches based on semi-supervision [33] such as Co-training [3] and Tri-training [45] which integrate seamlessly with our proposed framework. We plan to experiment with these variants of proxy labelling in future work.

Model Compression The larger size of the models exacerbates the challenge of deployment with limited resources [5, 34]. Multiple methods like quantization [44], pruning [10], distillation [14, 35] and weight sharing [13] are used to mitigate this issue. All these methods have shown varying degrees of success compared to the performance of the base model from which they are derived. [12] studied how a model can be used to label unlabelled data and make use of the model predictions for training using a combination of loss functions. In this paper, we consider BERT-based models as the teacher and the student.

3 METHODOLOGY

Our approach is focused on leveraging proxy labelling and knowledge distillation to build a highly accurate classifier with reduced cost of training and deployment. We trained a model using manually labeled data, and called it as the "teacher". We then passed the unlabelled data through this teacher model and used the predictions as the corresponding labels for the data. The manually labelled data combined with the proxy labelled data was then used to train a "student" model. In this paper, we experiment with two student model training strategies and choose the one that achieves the highest F1 score on our validation set. Figure 1 shows a high level system overview.

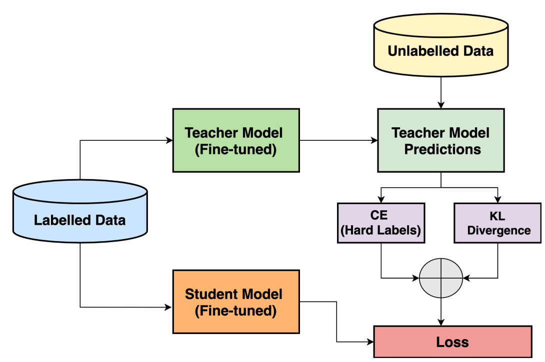

Figure 1: High level overview of the process of Knowledge Distillation using Proxy-Labelling

Cross entropy as the loss function with hard labels In this approach, we leveraged the teacher model to obtain weak labels for unlabelled data. For a given sample we assigned the most confident prediction from the teacher as its label. Concretely, the softmax output probabilities for each sample were converted into one hot encodings considering the class with highest probability as the true label.

KL Divergence as the loss function with soft labels Similar to the previous approach, we leveraged the teacher model to obtain weak labels on the unlabelled data. But instead of one hot encodings, we considered the probability distribution of the predictions as the labels for the samples. In other words, we performed semi-supervision while ensuring to replicate the teacher's behaviour when labelling the samples. Here, we made use of KL divergence [16] loss instead of Cross Entropy loss.

4 EXPERIMENTS

Dataset The initial labelled training data comprised of 41,539 customer transactions sampled from September to December (2019) time frame - only thsse transactions that had an associated transaction description provided by the customer. The samples were manually annotated by a team of three SMEs and mapped to one of the ten pre-defined categories. The list of categories considered are as follows: {'Food', 'Grocery', 'Package', 'Medicines', 'Household Items', 'Cigarettes', 'Clothes', 'Electronics', 'Keys', 'Documents/Books'}. Additionally, we considered 285,235 unlabelled customer transactions sampled from January to April (2020) for the proxy labelling experiments. For benchmarking the performance of different classification approaches, we used manually labeled 20,156 customer transactions from April (2020) to construct a validation dataset. This validation set, containing 20,156 samples, was not used for the proxy labelling experiments (or any other experiments).

Training the Teacher model In the first step, we trained multiple models using the *Train* dataset and validated on *Validation* dataset to identify the candidate teacher model for our Knowledge Distillation framework - we considered XgBoost [6, 28], BiLSTM [1, 9], ALBERT [18] and RoBERTa [22]. Table 2 shows the F1-scores for different models considered for this experiment. We observed that ALBERT and RoBERTa outperform BiLSTM and XgBoost, and RoBERTa outperformed ALBERT. Therefore, RoBERTa was chosen as the teacher model for the next set of experiments.

Model	F1 Score	Accuracy
XgBoost	0.60	63
BiLSTM	0.65	73
ALBERT	0.70	78
RoBERTa	**0.74**	**82**

Table 2: F1-scores and Accuracies for different classification models trained on Train dataset and validated on Validation dataset

Generate Weak Labels for Unlabelled Data In the second step, we leveraged the teacher model selected in the previous step to extract weakly labeled samples for the *Unlabelled* dataset to augment the training dataset. To reduce the probability of selecting mislabeled samples, we set an empirical threshold of 95% confidence in the label prediction as the criteria to accept a sample into the pool of training samples, thus, obtaining 93,820 additional training samples

Model	Parameters (in millions)
ALBERT(base)	**11**
distilBERT(base)	66
distilRoBERTa(base)	82
RoBERTa(base)	125
RoBERTa(large)	355

Table 3: Comparison of the number of parameters among different BERT-based models. ALBERT has the fewest number of parameters

Leverage Knowledge Distillation to train a Student model Due to productionization constraints on number of parameters in the model , ALBERT(base) with ˜11 million parameters was chosen as the student model from the set of SOTA models. The detailed comparison of the number of parameters can be found in Table 3. The label data generated from the Teacher model (as described in the previous sub-section) was further used to "teach" the student models. The student model R-ALBERT was first fine-tuned on the labelled *Train* dataset used to train the teacher and then further fine-tuned making use of the 3 strategies described in Section 3. This student model performed the best and even better than the teacher model on our *Validation* dataset. A Similar model behavior was observed in [8] and we plan to perform detailed analyses on this in future work.

Reproducibility We considered the 20Newsgroup [32] dataset to validate the reproducibility of our proposed approach on publicly available datasets.

Model	F1 Score	Accuracy
R-ALBERT - OHE	0.72	83
R-ALBERT - KL	**0.73**	**84**

Table 4: Comparison of F1-scores and accuracies on the internal benchmark using different approaches. We can notice that R-ALBERT with KL divergence performs better than R-ALBERT with OHE

Model 1	Model 2	Chi-square	p-value (<)
ALBERT	RoBERTa	2185.71	2.2e-16
R-ALBERT-KL	R-ALBERT-OHE	955.61	2.2e-16
R-ALBERT-KL	RoBERTa	955.61	2.2e-16

Table 5: Stuart-Maxwell Test shows that the performance improvements on accuracy with R-ALBERT-KL are statistically significant. here, the performance of *Model 1* is compared with that of *Model 2*

5 RESULTS

As shown in Table 4, the Student model tends to perform better than its base version (without the Teacher). We validate the statistical significance of the performance improvement using Stuart-Maxwell Test [24, 36]. As shown in Table 5, the performance improvement of R-ALBERT-KL over the base models are significant. Moreover, we observe that our approach achieved similar performance when compared to human annotated data, despite the change in data distributions and textual patterns. Also, from Table 6, we observe that the given method is reproducible on the 20Newsgroup dataset.

Model	F1-score	Accuracy (%)
ALBERT	0.63	65
R-ALBERT-KL	0.70	73
RoBERTa	0.88	87

Table 6: F1-scores on the 20Newsgroup dataset with 8,073 train samples, 7,037 weakly labeled samples (after 95% threshold) and 805 samples

Moreover, we observe that the distillation over RoBERTa's predictions gave an improvement of 8% when compared to fine-tuning only on the labelled dataset.

6 CONCLUSION

In this paper, we explored a generalised framework that combined proxy labelling with distillation on transformer based architectures. Here "students" can be made better with the help of the weak labels generated by a good "teacher". Given an industry setting focused on effective utilisation of resources, the proposed approach will pave the path for more research on reducing manual labelling and model size in the future. Furthermore, we also noted that the proxy labels generated gave comparable performance to human labelled data. Our current model is deployed and is handling category prediction at scale. An open area of research worth exploring is the multi-category classification within our framework.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*. 265–283.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 722–735.

[3] Avrim Blum and Tom Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* (Madison, Wisconsin, USA) *(COLT' 98)*. Association for Computing Machinery, New York, NY, USA, 92–100. https://doi.org/10.1145/279943.279962

[4] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit Dhillon. 2019. X-BERT: eXtreme Multi-label Text Classification with using Bidirectional Encoder Representations from Transformers. *arXiv preprint arXiv:1905.02331* (2019).

[5] Daoyuan Chen, Yaliang Li, Minghui Qiu, Zhen Wang, Bofang Li, Bolin Ding, Hongbo Deng, Jun Huang, Wei Lin, and Jingren Zhou. 2020. Adabert: Task-adaptive bert compression with differentiable neural architecture search. *arXiv preprint arXiv:2001.04246* (2020).

[6] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, and Yuan Tang. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2* (2015), 1–4.

[7] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. 2020. A Survey of Model Compression and Acceleration for Deep Neural Networks. arXiv:1710.09282 [cs.LG]

[8] Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kardas, Sylvain Gugger, and Jeremy Howard. 2019. Multifit: Efficient multi-lingual language model fine-tuning. *arXiv preprint arXiv:1909.04761* (2019).

[9] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with LSTM. (1999).

[10] Mitchell A Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing BERT: Studying the effects of weight pruning on transfer learning. *arXiv preprint arXiv:2002.08307* (2020).

[11] Swapnil Hingmire and Sutanu Chakraborti. 2014. Topic labeled text classification: A weakly supervised approach. In *ACM SIGIR Conference on Research and Development in Information Retrieval*.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[13] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. *arXiv preprint arXiv:1902.00751* (2019).

[14] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351* (2019).

[15] Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Unifying Question Answering, Text Classification, and Regression via Span Extraction. arXiv:1904.09286 [cs.CL]

[16] Solomon Kullback. 1997. *Information theory and statistics*. Courier Corporation.

[17] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. *arXiv preprint arXiv:1704.04683* (2017).

[18] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[19] Ximing Li and Bo Yang. 2018. A Pseudo Label based Dataless Naive Bayes Algorithm for Text. In *International Conference on Computational Linguistics*.

[20] Percy Liang. 2005. Semi-supervised learning for natural language. In *MASTER'S THESIS, MIT*.

[21] Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking Natural Language Understanding Services for building Conversational Agents. arXiv:1903.05566 [cs.CL]

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[23] Zhibin Lu, Pan Du, and Jian-Yun Nie. 2020. VGCN-BERT: Augmenting BERT with Graph Embedding for Text Classification. In *European Conference on Information Retrieval*. Springer, 369–382.

[24] Albert Ernest Maxwell. 1970. Comparing the classification of subjects by two independent judges. *The British Journal of Psychiatry* 116, 535 (1970), 651–655.

[25] David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective Self-Training for Parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. Association for Computational Linguistics, New York City, USA, 152–159. https://www.aclweb.org/anthology/N06-1020

[26] Dheeraj Mekala and Jingbo Shang. 2020. Contextualized Weak Supervision for Text Classification. In *Association for Computational Linguistics*.

[27] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. 2013. Learning with Noisy Labels. In *Neural Information Processing Systems*.

[28] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[29] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Association for Computational Linguistics*.

[30] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing*.

[31] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. In *Proceedings of the VLDB Endowment*.

[32] Jason Rennie. 2008. Newsgroups dataset, 2008.

[33] Sebastian Ruder. 2018. *An overview of proxy-label approaches for semi-supervised learning*. https://ruder.io/semi-supervised/index.html#fn19

[34] Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2020. Poor Man's BERT: Smaller and Faster Transformer Models. *arXiv preprint arXiv:2004.03844* (2020).

[35] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[36] Alan Stuart. 1955. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* 42, 3/4 (1955), 412–416.

[37] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. A Survey on Deep Transfer Learning. arXiv:1808.01974 [cs.LG]

[38] Tan Thongtan and Tanasanee Phienthrakul. 2019. Sentiment Classification Using Document Embeddings Trained with Cosine Similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Florence, Italy, 407–414. https://doi.org/10.18653/v1/P19-2057

[39] Ellen M. Voorhees and Donna Harman. 2000. Overview of the Sixth Text REtrieval Conference (TREC-6). *Inf. Process. Manage.* 36, 1 (Jan. 2000), 3–35. https://doi.org/10.1016/S0306-4573(99)00043-6

[40] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Empirical Methods in Natural Language Processing*.

[41] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* (2019), arXiv–1910.

[42] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification. arXiv:1905.00546 [cs.CV]

[43] David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *33rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Cambridge, Massachusetts, USA, 189–196. https://doi.org/10.3115/981658.981684

[44] Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8bert: Quantized 8bit bert. *arXiv preprint arXiv:1910.06188* (2019).

[45] Zhi-Hua Zhou and Ming Li. 2005. Tri-Training: Exploiting Unlabeled Data Using Three Classifiers. *IEEE Trans. on Knowl. and Data Eng.* 17, 11 (Nov. 2005), 1529–1541. https://doi.org/10.1109/TKDE.2005.186