Social Re-Identification Assisted RTO Detection for E-Commerce

Hitkul
hitkuli@iiitd.ac.in
IIIT Delhi
Delhi, India

Abinaya K
abinaya.k@flipkart.com
Flipkart
Bengaluru, India

Soham Saha
soham.saha@flipkart.com
Flipkart
Bengaluru, India

Satyajit Banerjee
satyajit.banerjee@flipkart.com
Flipkart
Bengaluru, India

Muthusamy Chelliah
muthusamy.c@flipkart.com
Flipkart
Bengaluru, India

Ponnurangam Kumaraguru
pk.guru@iiit.ac.in
IIIT Hyderabad
Hyderabad, India

ABSTRACT

E-commerce features like easy cancellations, returns, and refunds can be exploited by bad actors or uninformed customers, leading to revenue loss for organization. One such problem faced by e-commerce platforms is Return To Origin (RTO), where the user cancels an order while it is in transit for delivery. In such a scenario platform faces logistics and opportunity costs. Traditionally, models trained on historical trends are used to predict the propensity of an order becoming RTO. Sociology literature has highlighted clear correlations between socio-economic indicators and users' tendency to exploit systems to gain financial advantage. Social media profiles have information about location, education, and profession which have been shown to be an estimator of socio-economic condition. We believe combining social media data with e-commerce information can lead to improvements in a variety of tasks like RTO, recommendation, fraud detection, and credit modeling. In our proposed system, we find the public social profile of an e-commerce user and extract socio-economic features. Internal data fused with extracted social features are used to train a RTO order detection model. Our system demonstrates a performance improvement in RTO detection of 3.1% and 19.9% on precision and recall, respectively. Our system directly impacts the bottom line revenue and shows the applicability of social re-identification in e-commerce.

CCS CONCEPTS

• **Applied computing** → **E-commerce infrastructure**.

KEYWORDS

E-commerce, Social Re-identification, RTO Detection

ACM Reference Format:
Hitkul, Abinaya K, Soham Saha, Satyajit Banerjee, Muthusamy Chelliah, and Ponnurangam Kumaraguru. 2023. Social Re-Identification Assisted RTO Detection for E-Commerce. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3543873.3587620

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04…$15.00
https://doi.org/10.1145/3543873.3587620

1 INTRODUCTION

Over the last decade, e-commerce adoption has proliferated rapidly [19]. Such growth is fueled by convenience that e-commerce can provide over brick and mortar, e.g., large product selection, lower prices, same-day shipping, and hassle-free returns and cancellations. Though convenience features attract customers, they can sometimes cause significant business challenges; one such case is Return-to-Origin (RTO). RTO as depicted in Figure 1 is a scenario when a customer orders a product and then cancels while it is en route. RTO leads to two kinds of losses in a system:-

- **Logistical cost**: This is the cost of shipping the product till the point of cancellation in the supply chain and then returning it to the warehouse safely and restocking it.
- **Opportunity cost**: In the time while the product was ordered and canceled, this product unit became unavailable to order by another customer who would accept the delivery.

Though business accounts for potential revenue loss while offering functionality like RTO, an increased rate of RTO by uninformed customers or bad actors can cause unanticipated revenue losses totaling double-digit million dollars annually. Hence it becomes necessary to develop a real-time system that can predict the likelihood of the order being subjected to RTO at the time of checkout. Prediction of the model combined with other attributes like customer history, and available stock of the product can be used to initiate precautionary measures that can mitigate RTO risk. Naturally, the data used to build such a system would be, the historical pattern of RTOs at a user and product level. However, a system built on

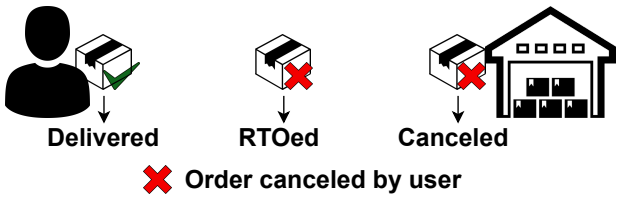

Figure 1: An order becomes Return to Origin (RTO) when the user cancels an order after it has been shipped from the source location.

these features is limited in its capability, especially for new users and product categories.

Literature has shown that socio-economic attributes of customer can be an indicator to identify the likelihood of a person being involved in activities like electricity theft [20], false insurance claims [26], or mortgage fraud [3]. Public social media profiles can be used to estimate socio-economic features [14]. Adding features from social media profiles has shown improved results in a variety of tasks, e.g., identifying transaction fraud [11], the credibility of online information [9], hate speech [4], and propensity to participate in risk-taking activities [13, 15]. Grounded in the aforementioned literature, we hypothesize that enriching historical data with publicly available social data of a consumer will lead to a performance improvement in RTO prediction.

The first step for our experiments is to re-identify the social profiles of a given user. The problem of social re-identification is studied widely [10, 17, 18, 24, 28]. Though most literature relates to retrieving matches between two social media platforms with a notable exception of [7], our task is slightly different, where we need to match profiles between a social and an e-commerce platform.

In this paper, given an e-commerce user, we find the relevant public social profile and show that the fusion of social information with historical trend data improves the performance of RTO prediction by 3.1% increase in precision and 19.9% increase in recall. Our work has direct implications for e-commerce platforms where a system like this can prevent loss of revenue. Additionally, our study demonstrates that combining social information with internal platform data can be a valuable tool for improving downstream tasks like RTO.

2 DATA AND SOCIAL RE-IDENTIFICATION

In this section, we first provide the details of our ground truth RTO dataset, followed by social re-identification candidate extraction (§ 2.2) and validation (§ 2.3) steps.

2.1 Ground Truth Data

We can extract ground truth from all past orders and their subsequent outcomes of the e-commerce platform. Orders are subjected to multiple internal models during checkout, which can induce unintended biases in the data. To prevent this, 5% of all orders are randomly set aside, as the *control set*, where no intervention is applied. Further, we extracted the cash on delivery orders from the control set, because we observed that orders with cash on delivery are more prone to RTO. All our experiments and benchmarking are performed against this set. Our experiments are performed on 6 months (November 2021 - April 2022) of data. First 5 months of data is used for training, and the following 1 month is used as a test set.

We ensure that our study design does not breach the privacy terms and conditions of our platform, or of the social media platforms used. As an extra layer of prevention, experiments shown in this work are performed only on users who explicitly decided to make their name and city locations[1] public on the platform. After all filtration, our final dataset includes 6,881 orders placed by 2,121 unique users. Out of all, 2,201 (32%) orders were RTOed.

[1]Used for social re-identification, see § 2.2

2.2 Potential Candidate Extraction

The initial step of user re-identification is to reduce the infinite search space of social profiles to a few candidate profiles for a given user. Querying social media platform's search engine using the *name* and *location* of a user has been shown to narrow the candidate pool effectively [7, 10]. For every unique user in our dataset, we create a search query of format *<user name> <city name>* and retrieve results from the social platform's search engines and a leading web search engine. Top 10 results of the query are used as candidate profiles.

We use a popular professional networking social media platform as a source of our social data; since, along with general information, such platforms have specific information that can reflect socio-economic indicators. Only data explicitly made public on the platform by the user is collected and used. Out of total 2,121 unique users, we found potential candidate profiles for 1,091 users.

2.3 Social Re-Identification

Literature shows that different social profile attributes like name, location, network, and language features can be used to find a match from candidate profiles [24]. Considering the asymmetry between e-commerce and social media platforms, all these attributes are not available on both the platforms. However, we are in a unique position to access various locations a user has ordered from in the past. [6, 7, 22, 27] showed that matching various location information in a user's profiles with candidate profiles can find correct matches with a high probability.

We perform candidate filtration using two attributes viz. names and locations. Firstly, any candidate profiles whose names do not match the source user are rejected. In the second step, given a source user u, we extract from the orders history a set L_u, defined as $\{l_u^1, l_u^2, ..., l_u^n\}$ where l_u^i is the i^{th} city u placed a order at. For each potential candidate profile of u, a similar location set L_u^c is defined as $\{l_c^1, l_c^2, ..., l_c^m\}$ where c denotes a candidate profile and l_c^i is a city location mentioned in c's social profile.

The Match score of candidate profile c with u (α_u^c) is defined as the ratio of location in social profiles also present in the source user location set. While calculating the intersection between the set of city names fuzzy matching was used to account for slight variation in spellings and syntax of city names. E.g., Delhi vs. New Delhi, or Bangalore vs. Bengaluru.

$$\alpha_u^c = \frac{|L_u^c \cap L_u|}{|L_u^c|} \tag{1}$$

A candidate profile is considered a match if α_u^c is above a predefined threshold θ. A user can be classified into three categories based on the number of matches received. 'No match' for users where no candidate profiles had a score above θ, an 'Exact match' where exactly 1 candidate profile had the matching score above θ, and 'Multiple matches' in which case we found more than one candidate profile who had match score above the threshold. Table 1 shows the percentage of users in each of three categories for different values of θ. Users in the 'No match' category were removed from the modeling step. In case of 'Multiple match', final feature value is obtained by averaging over all the matches. Results shown

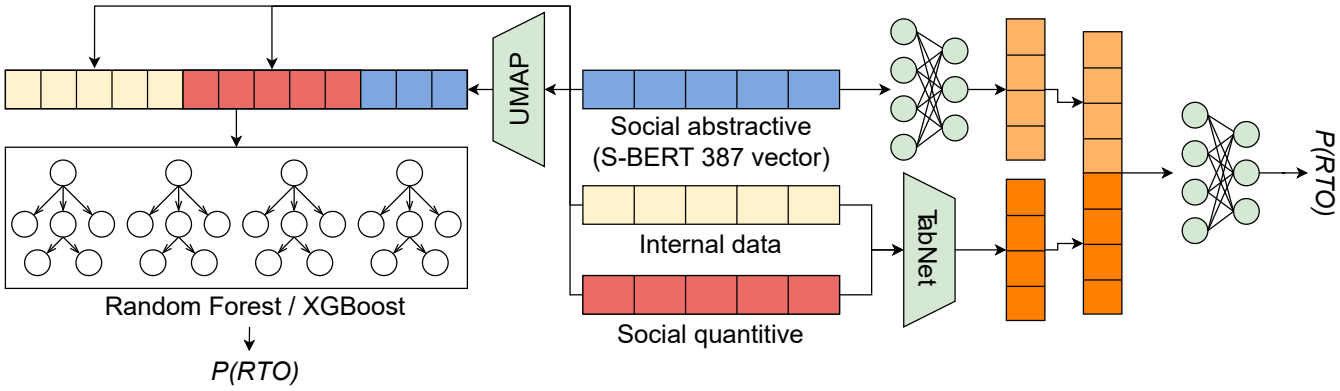

Figure 2: Our training architecture. In the case of tree-based models (on the left), all three feature sets are concatenated to form the input. While training deep learning models (on the right), tabular features are encoded via Tabnet and concatenated with S-BERT embeddings before being passed into a feed-forward neural network.

in this paper are calculated using $\theta = 0.6$, results for varying values of θ were consistent and are omitted due to lack of space.

3 RTO MODEL

We discuss the features used by our proposed model, the types of modeling techniques we experimented with, and the evaluation metrics used.

3.1 Features

We broadly divide the features used into three categories; 1) past trends, 2) social profile quantitative, and 3) social profile abstractive. The first category is derived from internal data, and the other two are extracted from social profiles.

Past trends: These features are derived from historical data. Each sample includes the ratio of RTO vs. total orders over the last 3 months and 1 year for the user, products in order, seller, and product category. Apart from this, location is also a robust socio-economic indicator; therefore, we extract the same trends for pin code, street, and city mentioned in the delivery address. Further, we observed a correlation between the RTO rate and the order time (specifically the hour and weekday). Hence hour of the day, weekday, and respective past trends are added to the feature list.

Social profile quantitative: As we identify social profiles for a user, we extract if the user is a student, number of jobs, number of educational degrees, and number of friends and followers. The count of jobs/degrees may not always be a good indicator of someone's professional position since some people spend a long time in the same jobs, whereas others often switch jobs. Pertaining to that, we add two features counting the total years a user has spent working and in education.

Social profile abstractive: We have extracted social features related to the quantity of experience and education of users. Research has shown that institutions of education and programs studied can significantly impact career success [23]. Similarly, two people with the same years of job experience can have widely different buying propensities based on what roles they are pursuing at which organizations. We hypothesize features capturing user's education institutes and job roles can assist in RTO prediction. Recently, contextual language models pretrained on large volumes of data, have captured and exploited complex relations well for downstream tasks [2, 5]. Following this, we extract the latest education institute, and the course pursued by a given user and pass this textual information via a pretrained Sentence-BERT [21] model to generate 387 dimension vectors. A similar vector is also created for the Job organization and designation the user had while placing the order.

Table 1: Results of social re-identification for varying values of matching threshold θ.

Match Threshold θ	Exact Match	Multiple Match	No Match
0.1	81.49	18.51	0.00
0.2	81.31	18.51	0.18
0.3	79.58	18.51	1.91
0.4	76.21	18.41	5.38
0.5	71.01	18.41	10.57
0.6	68.92	17.68	13.40
0.7	65.91	17.50	16.59
0.8	64.63	17.41	17.96
0.9	64.36	17.41	18.23
1.0	42.57	17.41	40.02

3.2 ML Modeling

Most of our data is tabular making tree-based ensemble methods like Random forest and XGBoost the default choice. Recently, attention-based architecture like Tabnet [1] has been proposed claiming to outperform traditional tree-based models. We present results on both types of models.

Figure 2 shows our training setup. In the tree-based models, 387 dimension vector obtained for job and education are decomposed to lower dimensions using UMAP [16] to prevent overfitting. The final

Table 2: RTO detection performance on the test set. Random forest performs the best. The addition of social features with past trend data increases goodness by 628 bps.

Model	Features	Precision (%)	Recall (%)	Goodness (bps)
Random Forest	Past Trends	85.7	40.3	1,005.7
	Past Trends + Social quantitative	85.7	50.4	1,305.6
	Past Trends + Social quantitative + Social abstractive	**88.8**	**60.2**	**1,633.7**
XGBoost	Past Trends	80.0	33.6	809.3
	Past Trends + Social quantitative	82.2	39.7	994.1
	Past Trends + Social quantitative + Social abstractive	86.8	44.5	1,129.4
TabNet	Past Trends	82.4	39.4	977.0
	Past Trends + Social quantitative	78.2	30.2	716.2
	Past Trends + Social quantitative + Social abstractive	64.2	15.1	320.0

dimension after decomposition is treated as a hyperparameter. Finally, decomposed vectors are added to the table of quantitative features as columns and fed into the model. When experimenting with deep learning-based models, tabular features are passed through Tabnet to generate a feature embedding. Generated embedding is concatenated with sentence-BERT embeddings (see § 3.1) and passed into a series of fully connected layers. All models are hyperparameter tuned using random search over; 4-fold cross-validation over the training data is used for parameter selection.

3.3 Evaluation

We use precision and recall to evaluate the performance of our models, but at a large scale, even very small improvements in model performance can lead to measurable revenue benefits. Additionally, traditional metrics may not always fit well in business discourse. Highlighting this, we define a metric named *Goodness* on which our models are evaluated.

Goodness : It reflects the improvement in recall performance. Defined in Equation 2, it calculates the reduction in the ratio of RTO orders after being evaluated by the model. Multiplication with 10^4 is performed to convert value into Basis Points (bps), this improves readability even while observing quantitatively small improvements. A higher value is better.

$$Goodness = \left(\frac{|P|}{|P| + |N|} - \frac{|P| - |P_{Pred\ and\ True}|}{|P| + |N| - |P_{Pred}|} \right) \times 10^4 \quad (2)$$

$$FPR = \frac{|P_{Pred}| - |P_{Pred\ and\ True}|}{|P_{Pred}|} \quad (3)$$

Here, P is set of RTO orders, and N is set of Delivered orders. P_{Pred} is set of orders predicted as RTO by a model, and $P_{Pred\ and\ True}$ is set of true positive RTO predictions.

Our aim is to choose a classification threshold that maximizes *Goodness* while maintaining the false positive rate (*FPR*) below a fixed value. [2] A high *FPR* means increased false interventions, reducing customer experience. Just like precision and recall, *Goodness* and *FPR* are a trade-off balance. High *Goodness* comes with an increase in *FPR*.

[2] *FPR* threshold is decided based on product requirement.

4 RESULTS

Table 2 shows performance of various RTO models on our test set. The random forest provides the overall best performance. As hypothesized, adding social features with past trends improves goodness by 300 bps, and adding contextual embeddings representing education and professional information improves the goodness further by 328 bps. This model has direct implications for improving the bottom-line revenue performance of an e-commerce organization.

Contrary to intuition, deep learning based models performed the worse. Comparative studies has shown that this behaviour is common in case of tabular data [8, 12, 25]. Studies compared the performance of Tabnet, and its contemporaries on a large variety of tabular data tasks, and concluded that these neural architectures do not perform consistently and are very sensitive to parameter tuning.

5 IMPLICATIONS AND ETHICAL CONSIDERATIONS

The primary goal of any e-commerce organization is to induce efficiency in the supply chain. Our work achieves this by detecting RTO orders with 20% better recall without degrading the precision. An argument can be made about our recall figure being low. More critical, as highlighted by our metric goodness, is the delta increase in performance because every 1% increase in the recall can add six figures to the bottom-line revenue for a large-scale organization. Further, our work demonstrates the effectiveness of including external data with internal e-commerce data to improve performance.

5.1 Ethics and Privacy

Our experiments make use of social information which requires special privacy consideration. The users explicitly made public all the social data used for our experiments. Further, only the users who decided to make name and location pair public on the platform are used. We ensure we comply with the privacy and data policies of all platforms. Other regions of the world may have different national privacy policies, which must be complied with while utilizing our approach.

5.2 Threat of Validity

Our organization operates in India, so all our experiments are conducted on data from the same region. Though the underlying hypothesis of socioeconomic factors assisting in predicting fraud is true for any region, specific user behavior and macroeconomics can affect the effectiveness of our approach and need to be evaluated. Though we still believe that the finding in our work is valuable, as the fundamental hypothesis is true universally, and we demonstrate results in one of the largest and fastest-growing e-commerce markets.[3]

6 CONCLUSION AND FUTURE WORK

Our study aims to improve the performance of a critical e-commerce problem RTO, where a user places an order and then cancels while the product is in transit, leading to logistics and opportunity cost. We hypothesize that fusing a users' social data with past RTO trend data can lead to improvements in performance. Towards this, we build a system to extract social profiles from popular professional networking social media platforms for a given user. Location-based matching is used to filter from the candidate matches. Finally, we extract quantitative and contextual features of matched profiles and demonstrate improvements of 3.1%, and 19.9% precision and recall, respectively, in the RTO detection task. Our work has direct implications for improving the bottom-line revenue of an e-commerce organization. Potential future directions of our work can be to experiment with transfer learning or multitask setup to see if social re-identification can help in other facets of e-commerce experience like review credibility or credit modeling. We would also like to extend our experiments to include data from a broader type of social media platforms.

ACKNOWLEDGMENTS

We thank Soumyasis Gun, Tanmay Sachan, Aravind Narayanan, Saksham Mrig, Vikas Goel, and K.V.M Naidu for their help and insights during this work.

REFERENCES

[1] Sercan Ö Arik and Tomas Pfister. 2021. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6679–6687.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[3] Andrew T Carswell and Douglas C Bachtel. 2009. Mortgage fraud: A risk factor analysis of affected communities. *Crime, law and social change* 52, 4 (2009), 347–364.

[4] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*. Springer, 693–696.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[6] Xing Gao, Wenli Ji, Yongjun Li, Yao Deng, and Wei Dong. 2018. User identification with spatio-temporal awareness across social networks. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1831–1834.

[7] Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. 2013. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web*. 447–458.

[8] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems* 34 (2021), 18932–18943.

[9] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International conference on social informatics*. Springer, 228–243.

[10] Paridhi Jain, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. @ i seek'fb. me' identifying users across multiple online social networks. In *Proceedings of the 22nd international conference on World Wide Web*. 1259–1268.

[11] Soheil Jamshidi and Mahmoud Reza Hashemi. 2012. An efficient data enrichment scheme for fraud detection using social network analysis. In *6th International Symposium on Telecommunications (IST)*. IEEE, 1082–1087.

[12] Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. 2021. Regularization is all you need: Simple neural nets can excel on tabular data. *arXiv preprint arXiv:2106.11189* (2021).

[13] Hemank Lamba, Shashank Srikanth, Dheeraj Reddy Pailla, Shwetanshu Singh, Karandeep Singh Juneja, and Ponnurangam Kumaraguru. 2020. Driving the last mile: Characterizing and understanding distracted driving posts on social networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 393–404.

[14] Vasileios Lampos, Nikolaos Aletras, Jens K Geyti, Bin Zou, and Ingemar J Cox. 2016. Inferring the socioeconomic status of social media users based on behaviour and language. In *European conference on information retrieval*. Springer, 689–695.

[15] Mark R Leary, Lydia R Tchividijian, and Brook E Kraxberger. 1994. Self-presentation can be hazardous to your health: impression management and health risk. *Health Psychology* 13, 6 (1994), 461.

[16] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[17] Ravita Mishra. 2019. Entity resolution in online multiple social networks (@ Facebook and LinkedIn). In *Emerging Technologies in Data Mining and Information Security*. Springer, 221–237.

[18] Ahmet Anıl Müngen, Esra Gündoğan, and Mehmet Kaya. 2021. Identifying multiple social network accounts belonging to the same users. *Social Network Analysis and Mining* 11, 1 (2021), 1–19.

[19] Shrey Nougarahiya, Gaurav Shetty, and Dheeraj Mandloi. 2021. A review of e–commerce in india: The past, present, and the future. *Research Review International Journal of Multidisciplinary* 6, 03 (2021), 12–22.

[20] Jonatas Pulz, Renan B Muller, Fabio Romero, André Meffe, Álvaro F Garcez Neto, and Aldo S Jesus. 2017. Fraud detection in low-voltage electricity consumers using socio-economic indicators and billing profile in smart grids. *CIRED-Open Access Proceedings Journal* 2017, 1 (2017), 2300–2303.

[21] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[22] Christopher Riederer, Yunsung Kim, Augustin Chaintreau, Nitish Korula, and Silvio Lattanzi. 2016. Linking users across domains with location data: Theory and validation. In *Proceedings of the 25th international conference on world wide web*. 707–719.

[23] Georgeanna FWB Robinson, Lisa S Schwartz, Linda A DiMeglio, Jasjit S Ahluwalia, and Janice L Gabrilove. 2016. Understanding career success and its contributing factors for clinical and translational investigators. *Academic medicine: journal of the Association of American Medical Colleges* 91, 4 (2016), 570.

[24] Kai Shu, Suhang Wang, Jiliang Tang, Reza Zafarani, and Huan Liu. 2017. User identity linkage across online social networks: A review. *Acm Sigkdd Explorations Newsletter* 18, 2 (2017), 5–17.

[25] Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion* 81 (2022), 84–90.

[26] Sharon Tennyson. 1997. Economic institutions and individual ethics: A study of consumer attitudes toward insurance fraud. *Journal of Economic Behavior & Organization* 32, 2 (1997), 247–265.

[27] Huandong Wang, Yong Li, Gang Wang, and Depeng Jin. 2021. Linking Multiple User Identities of Multiple Services from Massive Mobility Traces. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 4 (2021), 1–28.

[28] Reza Zafarani and Huan Liu. 2013. Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 41–49.

[3] https://www.business.com/articles/10-of-the-largest-ecommerce-markets-in-the-world-b/