



(51) International Patent Classification:

G06N 20/00 (2019.01) G06V 10/764 (2022.01)
G06N 3/09 (2023.01)

(21) International Application Number:

PCT/CA2024/050158

(22) International Filing Date:

08 February 2024 (08.02.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

18/394,697 22 December 2023 (22.12.2023) US

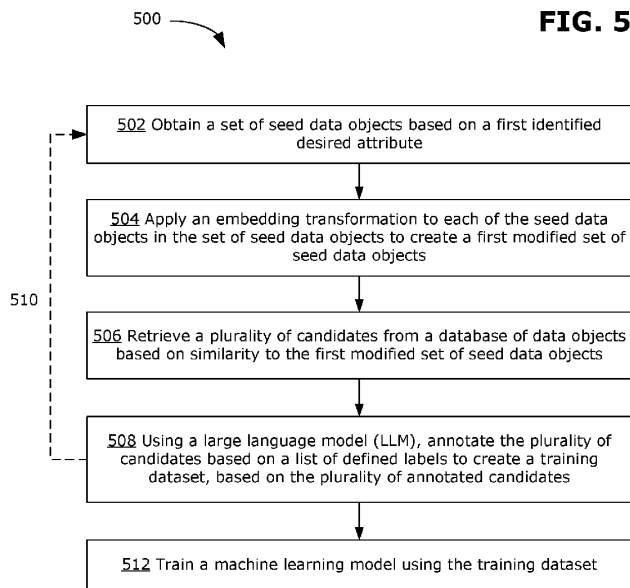
(71) Applicant: **SHOPIFY INC.** [CA/CA]; c/o Legal-IP Department, Shopify Inc., 151 O'Connor Street, Ground Floor, Ottawa, Ontario K2P 2L8 (CA).

(72) Inventors: **MATAMOROS, Javier Morales**; c/o Legal-IP Department, Shopify Inc., 151 O'Connor Street, Ground Floor, Ottawa, Ontario K2P 2L8 (CA). **MOHAN, Vandana**; c/o Legal-IP Department, Shopify Inc., 151 O'Connor Street, Ground Floor, Ottawa, Ontario K2P 2L8 (CA).

(74) Agent: **SMART & BIGGAR LP** et al.; P.O. Box 2999, Station D., 900-55 Metcalfe Street, Ottawa, Ontario K1P 5Y6 (CA).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH,

(54) Title: METHODS AND SYSTEMS FOR GENERATING LABELED TRAINING DATA USING A LLM



(57) Abstract: A method and apparatus is provided to automatically generate a training dataset and to train a machine learning (ML) model, such as an image classifier. A first set of seed data objects is obtained based on a desired attribute and may be transformed into a first modified set of seed data objects. The first modified set of seed data objects is processed, and a first plurality of candidates is retrieved based on a similarity to the first modified set of seed data objects. Using a large language model (LLM), the first plurality of candidates is annotated based on a list of defined labels to create a training dataset, for training the ML model. The disclosed method and apparatus may enable improved computational efficiency in generating targeted labeled training datasets.

[Continued on next page]

TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS,
ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

Published:

— *with international search report (Art. 21(3))*
— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*