

SYSTEM, METHOD, AND COMPUTER PROGRAM PRODUCT FOR DEBIASING EMBEDDING VECTORS OF MACHINE LEARNING MODELS

CROSS REFERENCE TO RELATED APPLICATION

[0001] The present application claims the benefit of U.S. Provisional Patent Application No. 63/167,737 filed March 30, 2021, the disclosure of which is hereby incorporated by reference in its entirety.

BACKGROUND

1. Field

[0002] The present disclosure relates to machine learning models and natural language processing. Particularly, but not exclusively, the present disclosure relates to debiasing embedding vectors of neural network machine learning models.

2. Technical Considerations

[0003] An embedding vector may include a relatively low-dimensional vector representation of relatively higher-dimensional vectors. For example, an embedding vector may represent a mapping of the higher-dimensional vector (e.g., representing a discrete and/or categorical variable, a high-dimensional data record, and/or the like) into a lower-dimensional space. In machine learning modeling, including neural networks, embedding vectors may include multi-dimensional, learned continuous vector representations of the higher dimensional vectors. Natural language processing (NLP) may map words to vector representations (e.g., embedding vectors, word embeddings, and/or the like). Embedding vectors with similar representations (e.g., relatively short distance between the two vectors) may be considered to be similar (e.g., have a similar meaning, value, context, and/or the like).

[0004] However, embedding vectors may reflect and/or amplify biases in data from which they are generated. For example, biased embedding vectors used in machine learning models may reduce the accuracy and efficiency of computer systems executing said machine learning models.

[0005] There is a need in the art for a technical solution to debias the geometry of embedding vectors to improve the accuracy, fairness, and efficiency of computer-driven, decision-making systems employing said embedding vectors.

SUMMARY

[0006] Additional features and advantages are realized through the techniques of the present disclosure. Other embodiments and aspects of the present disclosure are described in detail herein and are considered a part of the claimed disclosure.

[0007] According to non-limiting embodiments or aspects, provided is a computer-implemented method for debiasing embedding vectors of machine learning models. The method may include receiving, with at least one processor, a plurality of embedding vectors from a neural network model. The method may also include generating, with the at least one processor, two clusters of embedding vectors based on the plurality of embedding vectors. The two clusters may include a first cluster of embedding vectors expected to be biased in a first direction and a second cluster expected to be biased in a second direction. The method may further include determining, with the at least one processor, a first mean vector of the first cluster of embedding vectors and a second mean vector of the second cluster of embedding vectors. The method may further include modifying, with the at least one processor, the first mean vector by moving the first mean vector toward each embedding vector of the first cluster of embedding vectors to provide a plurality of first candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a first direction defined between the second mean vector and each first candidate vector to determine a bias score associated with each first candidate vector. The method may further include replacing, with the at least one processor, the first mean vector with a first candidate vector of the plurality of first candidate vectors based on the bias score of the first candidate vector. The method may further include modifying, with the at least one processor, the second mean vector by moving the second mean vector toward each embedding vector of the second cluster of embedding vectors to provide a plurality of second candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a second direction defined between the first mean vector and each second candidate vector to determine a bias score associated with each second candidate vector. The method may further include replacing, with the at least one processor, the second mean vector with a second candidate vector of the plurality of second candidate vectors based on the bias score of the second candidate vector. The method may further include repeating, with the at least one processor, the modifying of the first mean vector, the replacing of the first mean vector, the modifying of the second mean

vector, and the replacing of the second mean vector until an extremum of the bias score is reached. The method may further include, in response to reaching the extremum of the bias score, linearly projecting, with the at least one processor, each embedding vector of the plurality of embedding vectors along a direction defined between the first mean vector and the second mean vector to debias the plurality of embedding vectors.

[0008] In some non-limiting embodiments or aspects, the plurality of embedding vectors may be vector representations of word embeddings and the neural network model may include a natural language processing (NLP) model.

[0009] In some non-limiting embodiments or aspects, the plurality of embedding vectors may be vector representations of merchant identity embeddings generated from customer transaction data. The method may further include modifying, with the at least one processor, a machine learning model of a fraud monitoring system using the plurality of embedding vectors that have been debiased by linearly projecting each embedding vector of the plurality of embedding vectors using the direction defined between the first mean vector and the second mean vector.

[0010] In some non-limiting embodiments or aspects, the bias score may be calculated from a Word Embedding Association Test (WEAT). The bias score may also be calculated from an Embedding Coherence Test (ECT).

[0011] In some non-limiting embodiments or aspects, the method may further include, before modifying the first mean vector and the second mean vector, determining, with the at least one processor, an initial bias score by linearly projecting the embedding vectors of both the first cluster and the second cluster along an initial direction defined between the first mean vector and the second mean vector. The replacing of the first mean vector with the first candidate vector of the plurality of first candidate vectors may be based on maximizing a difference between the initial bias score and the bias score of the first candidate vector. The replacing of the second mean vector with the second candidate vector of the plurality of second candidate vectors may be based on maximizing a difference between the initial bias score and the bias score of the second candidate vector.

[0012] According to non-limiting embodiments or aspects, provided is a system for debiasing embedding vectors of machine learning models. The system may include a server including at least one processor. The server may be programmed or configured to receive a plurality of embedding vectors from a neural network model.

The server may be programmed or configured to generate two clusters of embedding vectors based on the plurality of embedding vectors, the two clusters comprising a first cluster of embedding vectors expected to be biased in a first direction and a second cluster expected to be biased in a second direction. The server may be programmed or configured to determine a first mean vector of the first cluster of embedding vectors and a second mean vector of the second cluster of embedding vectors. The server may be programmed or configured to modify the first mean vector by moving the first mean vector toward each embedding vector of the first cluster of embedding vectors to provide a plurality of first candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a first direction defined between the second mean vector and each first candidate vector to determine a bias score associated with each first candidate vector. The server may be programmed or configured to replace the first mean vector with a first candidate vector of the plurality of first candidate vectors based on the bias score of the first candidate vector. The server may be programmed or configured to modify the second mean vector by moving the second mean vector toward each embedding vector of the second cluster of embedding vectors to provide a plurality of second candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a second direction defined between the first mean vector and each second candidate vector to determine a bias score associated with each second candidate vector. The server may be programmed or configured to replace the second mean vector with a second candidate vector of the plurality of second candidate vectors based on the bias score of the second candidate vector. The server may be programmed or configured to repeat the modifying of the first mean vector, the replacing of the first mean vector, the modifying of the second mean vector, and the replacing of the second mean vector until an extremum of the bias score is reached. The server may be programmed or configured to, in response to reaching the extremum of the bias score, linearly project each embedding vector of the plurality of embedding vectors along a direction defined between the first mean vector and the second mean vector to debias the plurality of embedding vectors.

[0013] In some non-limiting embodiments or aspects, the plurality of embedding vectors may be vector representations of word embeddings and the neural network model comprises a natural language processing (NLP) model.

[0014] In some non-limiting embodiments or aspects, the plurality of embedding vectors may be vector representations of merchant identity embeddings generated from customer transaction data.

[0015] In some non-limiting embodiments or aspects, the server may be further programmed or configured to modify a machine learning model of a fraud monitoring system using the plurality of embedding vectors that have been debiased by linearly projecting each embedding vector of the plurality of embedding vectors using the direction defined between the first mean vector and the second mean vector.

[0016] In some non-limiting embodiments or aspects, the bias score may be calculated from a Word Embedding Association Test (WEAT).

[0017] In some non-limiting embodiments or aspects, the bias score may be calculated from an Embedding Coherence Test (ECT).

[0018] In some non-limiting embodiments or aspects, the server may be further programmed or configured to, before modifying the first mean vector and the second mean vector, determine an initial bias score by linearly projecting the embedding vectors of both the first cluster and the second cluster along an initial direction defined between the first mean vector and the second mean vector. The replacing of the first mean vector with the first candidate vector of the plurality of first candidate vectors may be based on maximizing a difference between the initial bias score and the bias score of the first candidate vector. The replacing of the second mean vector with the second candidate vector of the plurality of second candidate vectors may be based on maximizing a difference between the initial bias score and the bias score of the second candidate vector.

[0019] According to non-limiting embodiments or aspects, provided is a computer program product for debiasing embedding vectors of machine learning models. The computer program product may include at least one non-transitory computer-readable medium including program instructions that, when executed by at least one processor, cause the at least one processor to receive a plurality of embedding vectors from a neural network model. The instructions may cause the at least one processor to generate two clusters of embedding vectors based on the plurality of embedding vectors, the two clusters comprising a first cluster of embedding vectors expected to be biased in a first direction and a second cluster expected to be biased in a second direction. The instructions may cause the at least one processor to determine a first mean vector of the first cluster of embedding vectors and a second mean vector of the

second cluster of embedding vectors. The instructions may cause the at least one processor to modify the first mean vector by moving the first mean vector toward each embedding vector of the first cluster of embedding vectors to provide a plurality of first candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a first direction defined between the second mean vector and each first candidate vector to determine a bias score associated with each first candidate vector. The instructions may cause the at least one processor to replace the first mean vector with a first candidate vector of the plurality of first candidate vectors based on the bias score of the first candidate vector. The instructions may cause the at least one processor to modify the second mean vector by moving the second mean vector toward each embedding vector of the second cluster of embedding vectors to provide a plurality of second candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a second direction defined between the first mean vector and each second candidate vector to determine a bias score associated with each second candidate vector. The instructions may cause the at least one processor to replace the second mean vector with a second candidate vector of the plurality of second candidate vectors based on the bias score of the second candidate vector. The instructions may cause the at least one processor to repeat the modifying of the first mean vector, the replacing of the first mean vector, the modifying of the second mean vector, and the replacing of the second mean vector until an extremum of the bias score is reached. The instructions may cause the at least one processor to, in response to reaching the extremum of the bias score, linearly project each embedding vector of the plurality of embedding vectors along a direction defined between the first mean vector and the second mean vector to debias the plurality of embedding vectors.

[0020] In some non-limiting embodiments or aspects, the plurality of embedding vectors may be vector representations of word embeddings and the neural network model comprises a natural language processing (NLP) model.

[0021] In some non-limiting embodiments or aspects, the plurality of embedding vectors may be vector representations of merchant identity embeddings generated from customer transaction data.

[0022] In some non-limiting embodiments or aspects, the instructions, when executed by at least one processor, may further cause the at least one processor to modify a machine learning model of a fraud monitoring system using the plurality of

embedding vectors that have been debiased by linearly projecting each embedding vector of the plurality of embedding vectors using the direction defined between the first mean vector and the second mean vector.

[0023] In some non-limiting embodiments or aspects, the bias score may be calculated from a Word Embedding Association Test (WEAT).

[0024] In some non-limiting embodiments or aspects, the bias score may be calculated from an Embedding Coherence Test (ECT).

[0025] In some non-limiting embodiments or aspects, the instructions, when executed by at least one processor, may further cause the at least one processor to, before modifying the first mean vector and the second mean vector, determine an initial bias score by linearly projecting the embedding vectors of both the first cluster and the second cluster along an initial direction defined between the first mean vector and the second mean vector. The replacing of the first mean vector with the first candidate vector of the plurality of first candidate vectors may be based on maximizing a difference between the initial bias score and the bias score of the first candidate vector. The replacing of the second mean vector with the second candidate vector of the plurality of second candidate vectors may be based on maximizing a difference between the initial bias score and the bias score of the second candidate vector.

[0026] Other non-limiting embodiments or aspects of the present disclosure will be set forth in the following numbered clauses:

[0027] Clause 1: A computer-implemented method comprising: receiving, with at least one processor, a plurality of embedding vectors from a neural network model; generating, with the at least one processor, two clusters of embedding vectors based on the plurality of embedding vectors, the two clusters comprising a first cluster of embedding vectors expected to be biased in a first direction and a second cluster expected to be biased in a second direction; determining, with the at least one processor, a first mean vector of the first cluster of embedding vectors and a second mean vector of the second cluster of embedding vectors; modifying, with the at least one processor, the first mean vector by moving the first mean vector toward each embedding vector of the first cluster of embedding vectors to provide a plurality of first candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a first direction defined between the second mean vector and each first candidate vector to determine a bias score associated with each first candidate vector; replacing, with the at least one processor, the first mean vector with

a first candidate vector of the plurality of first candidate vectors based on the bias score of the first candidate vector; modifying, with the at least one processor, the second mean vector by moving the second mean vector toward each embedding vector of the second cluster of embedding vectors to provide a plurality of second candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a second direction defined between the first mean vector and each second candidate vector to determine a bias score associated with each second candidate vector; replacing, with the at least one processor, the second mean vector with a second candidate vector of the plurality of second candidate vectors based on the bias score of the second candidate vector; repeating, with the at least one processor, the modifying of the first mean vector, the replacing of the first mean vector, the modifying of the second mean vector, and the replacing of the second mean vector until an extremum of the bias score is reached; and, in response to reaching the extremum of the bias score, linearly projecting, with the at least one processor, each embedding vector of the plurality of embedding vectors along a direction defined between the first mean vector and the second mean vector to debias the plurality of embedding vectors.

[0028] Clause 2: The computer-implemented method of clause 1, wherein the plurality of embedding vectors are vector representations of word embeddings and the neural network model comprises a natural language processing (NLP) model.

[0029] Clause 3: The computer-implemented method of clause 1 or 2, wherein the plurality of embedding vectors are vector representations of merchant identity embeddings generated from customer transaction data.

[0030] Clause 4: The computer-implemented method of any one of clauses 1-3, further comprising modifying, with the at least one processor, a machine learning model of a fraud monitoring system using the plurality of embedding vectors that have been debiased by linearly projecting each embedding vector of the plurality of embedding vectors using the direction defined between the first mean vector and the second mean vector.

[0031] Clause 5: The computer-implemented method of any one of clauses 1-4, wherein the bias score is calculated from a Word Embedding Association Test (WEAT).

[0032] Clause 6: The computer-implemented method of any one of clauses 1-5, wherein the bias score is calculated from an Embedding Coherence Test (ECT).

[0033] Clause 7: The computer-implemented method of any one of clauses 1-6, further comprising, before modifying the first mean vector and the second mean vector, determining, with the at least one processor, an initial bias score by linearly projecting the embedding vectors of both the first cluster and the second cluster along an initial direction defined between the first mean vector and the second mean vector, wherein: the replacing of the first mean vector with the first candidate vector of the plurality of first candidate vectors is based on maximizing a difference between the initial bias score and the bias score of the first candidate vector; and the replacing of the second mean vector with the second candidate vector of the plurality of second candidate vectors is based on maximizing a difference between the initial bias score and the bias score of the second candidate vector.

[0034] Clause 8: A system comprising a server comprising at least one processor, the server programmed or configured to: receive a plurality of embedding vectors from a neural network model; generate two clusters of embedding vectors based on the plurality of embedding vectors, the two clusters comprising a first cluster of embedding vectors expected to be biased in a first direction and a second cluster expected to be biased in a second direction; determine a first mean vector of the first cluster of embedding vectors and a second mean vector of the second cluster of embedding vectors; modify the first mean vector by moving the first mean vector toward each embedding vector of the first cluster of embedding vectors to provide a plurality of first candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a first direction defined between the second mean vector and each first candidate vector to determine a bias score associated with each first candidate vector; replace the first mean vector with a first candidate vector of the plurality of first candidate vectors based on the bias score of the first candidate vector; modify the second mean vector by moving the second mean vector toward each embedding vector of the second cluster of embedding vectors to provide a plurality of second candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a second direction defined between the first mean vector and each second candidate vector to determine a bias score associated with each second candidate vector; replace the second mean vector with a second candidate vector of the plurality of second candidate vectors based on the bias score of the second candidate vector; repeat the modifying of the first mean vector, the replacing of the first mean vector, the modifying of the second mean vector, and the

replacing of the second mean vector until an extremum of the bias score is reached; and in response to reaching the extremum of the bias score, linearly project each embedding vector of the plurality of embedding vectors along a direction defined between the first mean vector and the second mean vector to debias the plurality of embedding vectors.

[0035] Clause 9: The system of clause 8, wherein the plurality of embedding vectors are vector representations of word embeddings and the neural network model comprises a natural language processing (NLP) model.

[0036] Clause 10: The system of clause 8 or 9, wherein the plurality of embedding vectors are vector representations of merchant identity embeddings generated from customer transaction data.

[0037] Clause 11: The system of any one of clauses 8-10, wherein the server is further programmed or configured to modify a machine learning model of a fraud monitoring system using the plurality of embedding vectors that have been debiased by linearly projecting each embedding vector of the plurality of embedding vectors using the direction defined between the first mean vector and the second mean vector.

[0038] Clause 12: The system of any one of clauses 8-11, wherein the bias score is calculated from a Word Embedding Association Test (WEAT).

[0039] Clause 13: The system of any one of clauses 8-12, wherein the bias score is calculated from an Embedding Coherence Test (ECT).

[0040] Clause 14: The system of any one of clauses 8-13, wherein the server is further programmed or configured to, before modifying the first mean vector and the second mean vector, determine an initial bias score by linearly projecting the embedding vectors of both the first cluster and the second cluster along an initial direction defined between the first mean vector and the second mean vector, and wherein: the replacing of the first mean vector with the first candidate vector of the plurality of first candidate vectors is based on maximizing a difference between the initial bias score and the bias score of the first candidate vector; and the replacing of the second mean vector with the second candidate vector of the plurality of second candidate vectors is based on maximizing a difference between the initial bias score and the bias score of the second candidate vector.

[0041] Clause 15: A computer program product comprising at least one non-transitory computer-readable medium including program instructions that, when executed by at least one processor, cause the at least one processor to: receive a

plurality of embedding vectors from a neural network model; generate two clusters of embedding vectors based on the plurality of embedding vectors, the two clusters comprising a first cluster of embedding vectors expected to be biased in a first direction and a second cluster expected to be biased in a second direction; determine a first mean vector of the first cluster of embedding vectors and a second mean vector of the second cluster of embedding vectors; modify the first mean vector by moving the first mean vector toward each embedding vector of the first cluster of embedding vectors to provide a plurality of first candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a first direction defined between the second mean vector and each first candidate vector to determine a bias score associated with each first candidate vector; replace the first mean vector with a first candidate vector of the plurality of first candidate vectors based on the bias score of the first candidate vector; modify the second mean vector by moving the second mean vector toward each embedding vector of the second cluster of embedding vectors to provide a plurality of second candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a second direction defined between the first mean vector and each second candidate vector to determine a bias score associated with each second candidate vector; replace the second mean vector with a second candidate vector of the plurality of second candidate vectors based on the bias score of the second candidate vector; repeat the modifying of the first mean vector, the replacing of the first mean vector, the modifying of the second mean vector, and the replacing of the second mean vector until an extremum of the bias score is reached; and in response to reaching the extremum of the bias score, linearly project each embedding vector of the plurality of embedding vectors along a direction defined between the first mean vector and the second mean vector to debias the plurality of embedding vectors.

[0042] Clause 16: The computer program product of clause 15, wherein the plurality of embedding vectors are vector representations of word embeddings and the neural network model comprises a natural language processing (NLP) model.

[0043] Clause 17: The computer program product of clause 15 or 16, wherein the plurality of embedding vectors are vector representations of merchant identity embeddings generated from customer transaction data, and wherein the instructions, when executed by at least one processor, further cause the at least one processor to modify a machine learning model of a fraud monitoring system using the plurality of

embedding vectors that have been debiased by linearly projecting each embedding vector of the plurality of embedding vectors using the direction defined between the first mean vector and the second mean vector.

[0044] Clause 18: The computer program product of any one of clauses 15-17, wherein the bias score is calculated from a Word Embedding Association Test (WEAT).

[0045] Clause 19: The computer program product of any one of clauses 15-18, wherein the bias score is calculated from an Embedding Coherence Test (ECT).

[0046] Clause 20: The computer program product of any one of clauses 15-19, wherein the instructions, when executed by at least one processor, further cause the at least one processor to, before modifying the first mean vector and the second mean vector, determine an initial bias score by linearly projecting the embedding vectors of both the first cluster and the second cluster along an initial direction defined between the first mean vector and the second mean vector, and wherein: the replacing of the first mean vector with the first candidate vector of the plurality of first candidate vectors is based on maximizing a difference between the initial bias score and the bias score of the first candidate vector; and the replacing of the second mean vector with the second candidate vector of the plurality of second candidate vectors is based on maximizing a difference between the initial bias score and the bias score of the second candidate vector.

[0047] These and other features and characteristics of the present disclosure, as well as the methods of operation and functions of the related elements of structures and the combination of parts and economies of manufacture, will become more apparent upon consideration of the following description and the appended claims with reference to the accompanying drawings, all of which form a part of this specification, wherein like reference numerals designate corresponding parts in the various figures. It is to be expressly understood, however, that the drawings are for the purpose of illustration and description only and are not intended as a definition of the limits of the present disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

[0048] Additional advantages and details of the present disclosure are explained in greater detail below with reference to the exemplary embodiments or aspects that are illustrated in the accompanying schematic figures, in which:

[0049] FIG. 1 is a diagram of a system for debiasing embedding vectors of machine learning models, according to some non-limiting embodiments or aspects;

[0050] FIG. 2 illustrates example components of a device or system used in connection with non-limiting embodiments or aspects;

[0051] FIG. 3 is a flow diagram of a method for debiasing embedding vectors of machine learning models, according to some non-limiting embodiments or aspects;

[0052] FIGS. 4A-4D are diagrams of an exemplary implementation of the method shown in FIG. 3, according to some non-limiting embodiments or aspects;

[0053] FIG. 5 is a diagram of an exemplary implementation of the method shown in FIG. 3, according to some non-limiting embodiments or aspects;

[0054] FIGS. 6A-6C are graphical visualizations of exemplary vectors of an exemplary implementation of the process shown in FIG. 3, according to some non-limiting embodiments or aspects; and

[0055] FIG. 7 is a diagram of an exemplary environment in which methods, systems, and/or computer program products, described herein, may be implemented, according to some non-limiting embodiments or aspects.

[0056] It should be appreciated by those skilled in the art that any block diagrams herein represent conceptual views of illustrative systems embodying the principles of the present subject matter. Similarly, it may be appreciated that any flow charts, flow diagrams, state transition diagrams, pseudo code, and the like represent various processes which may be substantially represented in computer readable medium and executed by a computer or processor, whether or not such computer or processor is explicitly shown.

DETAILED DESCRIPTION

[0057] In the present document, the word "exemplary" is used herein to mean "serving as an example, instance, or illustration." Any embodiment or implementation of the present subject matter described herein as "exemplary" is not necessarily to be construed as preferred or advantageous over other embodiments or aspects.

[0058] The terms "comprises", "includes" "comprising", "including" or any other variations thereof, are intended to cover a non-exclusive inclusion, such that a setup, device, or method that comprises a list of components or steps does not include only those components or steps but may include other components or steps not expressly listed or inherent to such setup, device, or method. In other words, one or more

elements in a system or apparatus proceeded by “comprises... a” or “includes... a” does not, without more constraints, preclude the existence of other elements or additional elements in the system or apparatus.

[0059] For purposes of the description hereinafter, the terms “upper”, “lower”, “right”, “left”, “vertical”, “horizontal”, “top”, “bottom”, “lateral”, “longitudinal,” and derivatives thereof shall relate to non-limiting embodiments or aspects as they are oriented in the drawing figures. However, it is to be understood that non-limiting embodiments or aspects may assume various alternative variations and step sequences, except where expressly specified to the contrary. It is also to be understood that the specific devices and processes illustrated in the attached drawings, and described in the following specification, are simply exemplary embodiments or aspects. Hence, specific dimensions and other physical characteristics related to the embodiments or aspects disclosed herein are not to be considered as limiting.

[0060] No aspect, component, element, structure, act, step, function, instruction, and/or the like used herein should be construed as critical or essential unless explicitly described as such. Also, as used herein, the articles “a” and “an” are intended to include one or more items and may be used interchangeably with “one or more” and “at least one.” Furthermore, as used herein, the term “set” is intended to include one or more items (e.g., related items, unrelated items, a combination of related and unrelated items, etc.) and may be used interchangeably with “one or more” or “at least one.” Where only one item is intended, the term “one” or similar language is used. Also, as used herein, the terms “has,” “have,” “having,” or the like are intended to be open-ended terms. Further, the phrase “based on” is intended to mean “based at least partially on” unless explicitly stated otherwise.

[0061] Some non-limiting embodiments or aspects are described herein in connection with thresholds. As used herein, satisfying a threshold may refer to a value being greater than the threshold, more than the threshold, higher than the threshold, greater than or equal to the threshold, less than the threshold, fewer than the threshold, lower than the threshold, less than or equal to the threshold, equal to the threshold, and/or the like.

[0062] As used herein, the terms “communication” and “communicate” may refer to the reception, receipt, transmission, transfer, provision, and/or the like, of information (e.g., data, signals, messages, instructions, commands, and/or the like).

For one unit (e.g., a device, a system, a component of a device or system, combinations thereof, and/or the like) to be in communication with another unit means that the one unit is able to directly or indirectly receive information from and/or transmit information to the other unit. This may refer to a direct or indirect connection (e.g., a direct communication connection, an indirect communication connection, and/or the like) that is wired and/or wireless in nature. Additionally, two units may be in communication with each other even though the information transmitted may be modified, processed, relayed, and/or routed between the first and second unit. For example, a first unit may be in communication with a second unit even though the first unit passively receives information and does not actively transmit information to the second unit. As another example, a first unit may be in communication with a second unit if at least one intermediary unit (e.g., a third unit located between the first unit and the second unit) processes information received from the first unit and communicates the processed information to the second unit. In some non-limiting embodiments or aspects, a message may refer to a network packet (e.g., a data packet, and/or the like) that includes data. Any known electronic communication protocols and/or algorithms may be used, such as, for example, Transmission Control Protocol/Internet Protocol (TCP/IP) (including Hypertext Transfer Protocol (HTTP) and other protocols), wireless local area network (WLAN) (including 802.11 and other radio frequency-based protocols and methods), analog transmissions, cellular networks (e.g., Global System for Mobile Communications (GSM®), Code-Division Multiple Access (CDMA), Long-Term Evolution (LTE®), Worldwide Interoperability for Microwave Access (WiMAX®), a third generation (3G) network, a fourth generation (4G) network, a fifth generation (5G) network, etc.), and/or the like. It will be appreciated that numerous other arrangements are possible.

[0063] As used herein, the term “payment device” may refer to a portable financial device, an electronic payment device, a payment card (e.g., a credit or debit card), a gift card, a smartcard, smart media, a payroll card, a healthcare card, a wristband, a machine-readable medium containing account information, a keychain device or fob, a radio-frequency identification (RFID) transponder, a retailer discount or loyalty card, a cellular phone, an electronic wallet mobile application, a personal digital assistant (PDA), a pager, a security card, a computer, an access card, a wireless terminal, a transponder, and/or the like. In some non-limiting embodiments or aspects, the

payment device may include volatile or non-volatile memory to store information (e.g., an account identifier, a name of the account holder, and/or the like).

[0064] As used herein, the term “transaction service provider” may refer to an entity that receives transaction authorization requests from merchants or other entities and provides guarantees of payment, in some cases through an agreement between the transaction service provider and an issuer institution. For example, a transaction service provider may include a payment network such as Visa® or any other entity that processes transactions. The term “transaction processing system” may refer to one or more computer systems operated by or on behalf of a transaction service provider, such as a transaction processing server executing one or more software applications, a token service executing one or more software applications, and/or the like. A transaction processing server may include one or more processors and, in some non-limiting embodiments or aspects, may be operated by or on behalf of a transaction service provider.

[0065] As used herein, the term “payment gateway” may refer to an entity and/or a payment processing system operated by or on behalf of such an entity (e.g., a merchant service provider, a payment service provider, a payment facilitator, a payment facilitator that contracts with an acquirer, a payment aggregator, and/or the like), which provides payment services (e.g., transaction service provider payment services, payment processing services, and/or the like) to one or more merchants. The payment services may be associated with the use of portable financial devices managed by a transaction service provider. As used herein, the term “payment gateway system” may refer to one or more computer systems, computer devices, servers, groups of servers, and/or the like operated by or on behalf of a payment gateway.

[0066] As used herein, the term “issuer institution” may refer to one or more entities, such as a bank, that provide accounts to customers for conducting transactions (e.g., payment transactions), such as initiating credit and/or debit payments. For example, an issuer institution may provide an account identifier, such as a primary account number (PAN), to a customer that uniquely identifies one or more accounts associated with that customer. The account identifier may be embodied on a payment device, such as a physical payment instrument, e.g., a payment card, and/or may be electronic and used for electronic payments. The term “issuer system” refers to one or more computer systems operated by or on behalf of an issuer institution, such as a server

computer executing one or more software applications. For example, an issuer system may include one or more authorization servers for authorizing a transaction.

[0067] As used herein, the term “acquirer institution” may refer to an entity licensed and/or approved by the transaction service provider to originate transactions (e.g., payment transactions) using a payment device associated with the transaction service provider. The transactions the acquirer institution may originate may include payment transactions (e.g., purchases, original credit transactions (OCTs), account funding transactions (AFTs), and/or the like). In some non-limiting embodiments or aspects, an acquirer institution may be a bank. As used herein, the term “acquirer system” may refer to one or more computer systems, computer devices, software applications, and/or the like operated by or on behalf of an acquirer institution.

[0068] As used herein, the terms “authenticating system” and “authentication system” may refer to one or more computing devices that authenticate a user and/or an account, such as, but not limited to, a transaction processing system, merchant system, issuer system, payment gateway, a third-party authenticating service, and/or the like.

[0069] As used herein, the terms “request,” “response,” “request message,” and “response message” may refer to one or more messages, data packets, signals, and/or data structures used to communicate data between two or more components or units.

[0070] As used herein, the term “account identifier” may include one or more PANs, tokens, or other identifiers associated with a customer account. The term “token” may refer to an identifier that is used as a substitute or replacement identifier for an original account identifier, such as a PAN. Account identifiers may be alphanumeric or any combination of characters and/or symbols. Tokens may be associated with a PAN or other original account identifier in one or more data structures (e.g., one or more databases and/or the like) such that they may be used to conduct a transaction without directly using the original account identifier. In some examples, an original account identifier, such as a PAN, may be associated with a plurality of tokens for different individuals or purposes.

[0071] As used herein, the term “merchant” may refer to one or more entities (e.g., operators of retail businesses) that provide goods and/or services, and/or access to goods and/or services, to a user (e.g., a customer, a consumer, and/or the like) based on a transaction, such as a payment transaction. As used herein, “merchant system”

may refer to one or more computer systems operated by or on behalf of a merchant, such as a server executing one or more software applications. As used herein, the term “product” may refer to one or more goods and/or services offered by a merchant.

[0072] As used herein, a “point-of-sale (POS) device” may refer to one or more devices, which may be used by a merchant to conduct a transaction (e.g., a payment transaction) and/or process a transaction. For example, a POS device may include one or more client devices. Additionally or alternatively, a POS device may include peripheral devices, card readers, scanning devices (e.g., code scanners), Bluetooth® communication receivers, near-field communication (NFC) receivers, radio frequency identification (RFID) receivers, and/or other contactless transceivers or receivers, contact-based receivers, payment terminals, and/or the like. As used herein, a “point-of-sale (POS) system” may refer to one or more client devices and/or peripheral devices used by a merchant to conduct a transaction. For example, a POS system may include one or more POS devices and/or other like devices that may be used to conduct a payment transaction. In some non-limiting embodiments or aspects, a POS system (e.g., a merchant POS system) may include one or more server computers programmed or configured to process online payment transactions through webpages, mobile applications, and/or the like.

[0073] As used herein, the terms “client” and “client device” may refer to one or more client-side devices or systems (e.g., remote from a transaction service provider) used to initiate or facilitate a transaction (e.g., a payment transaction). As an example, a “client device” may refer to one or more POS devices used by a merchant, one or more acquirer host computers used by an acquirer, one or more mobile devices used by a user, and/or the like. In some non-limiting embodiments or aspects, a client device may be an electronic device configured to communicate with one or more networks and initiate or facilitate transactions. For example, a client device may include one or more computers, portable computers, laptop computers, tablet computers, mobile devices, cellular phones, wearable devices (e.g., watches, glasses, lenses, clothing, and/or the like), PDAs, and/or the like. Moreover, a “client” may also refer to an entity (e.g., a merchant, an acquirer, and/or the like) that owns, utilizes, and/or operates a client device for initiating transactions (e.g., for initiating transactions with a transaction service provider).

[0074] As used herein, the term “computing device” may refer to one or more electronic devices that are configured to directly or indirectly communicate with or over

one or more networks. A computing device may be a mobile device, a desktop computer, and/or any other like device. Furthermore, the term “computer” may refer to any computing device that includes the necessary components to receive, process, and output data, and normally includes a display, a processor, a memory, an input device, and a network interface. As used herein, the term “server” may refer to or include one or more processors or computers, storage devices, or similar computer arrangements that are operated by or facilitate communication and/or processing in a network environment, such as the Internet, although it will be appreciated that communication may be facilitated over one or more public or private network environments and that various other arrangements are possible. Further, multiple computers, e.g., servers, or other computerized devices, such as POS devices, directly or indirectly communicating in the network environment may constitute a “system,” such as a merchant’s POS system.

[0075] The term “processor,” as used herein, may represent any type of processing unit, such as a single processor having one or more cores, one or more cores of one or more processors, multiple processors each having one or more cores, and/or other arrangements and combinations of processing units.

[0076] As used herein, the term “system” may refer to one or more computing devices or combinations of computing devices (e.g., processors, servers, client devices, software applications, components of such, and/or the like). Reference to “a device,” “a server,” “a processor,” and/or the like, as used herein, may refer to a previously-recited device, server, or processor that is recited as performing a previous step or function, a different server or processor, and/or a combination of servers and/or processors. For example, as used in the specification and the claims, a first server or a first processor that is recited as performing a first step or a first function may refer to the same or different server or the same or different processor recited as performing a second step or a second function.

[0077] As used herein, “embedding vector” may refer to a relatively low-dimensional vector representation of relatively higher-dimensional vectors. For example, an embedding vector may represent a mapping of the higher-dimensional vector (e.g., representing a discrete and/or categorical variable, a high-dimensional data record, and/or the like) into a lower-dimensional space. In some non-limiting embodiments or aspects, an embedding vector may include a vector of continuous elements (e.g., numbers) used to represent a discrete variable (e.g., a categorical

variable and/or the like) for use in machine learning models (e.g., deep learning models, neural network models, and/or the like). For example, embedding vectors may include multi-dimensional, learned, continuous vector representations of discrete variables that may reduce the dimensionality of categorical variables and meaningfully represent categories in the transformed space. In some non-limiting embodiments or aspects, embedding vectors may simplify machine learning on large inputs, such as sparse vectors representing words. For example, an embedding vector may capture semantics of an input (e.g., a word from a corpus of documents) by placing semantically similar inputs close together in the embedding space. In some non-limiting embodiments or aspects, an embedding vector may be learned and reused across models. In some non-limiting embodiments or aspects, an embedding vector may include a vector representation of a data string (e.g., a word, a merchant identifier (e.g., name, identification number, etc.) based on said string's distributions in a corpus of data (e.g., a set of documents (e.g., books and/or the like), consumer transaction data (e.g., a series of merchant identifiers that each consumer account of a set of accounts transacted with and/or the like), etc.).

[0078] Non-limiting embodiments or aspects of the present disclosure are directed to a system, method, and computer program product for debiasing embedding vectors of neural network machine learning models. It will be appreciated that described systems and methods improve the accuracy and efficiency of computer-driven, decision-making systems by reducing and/or removing bias from embedding vectors in machine learning models. Because embedding vectors may similarly represent input variables having similar semantic values, and because bias may be amplified in embedding vectors, the removal of bias allows a machine learning model to detect semantic similarity resulting from other feature dimensions besides a source of bias. For example, semantic similarities that otherwise would have been detected but are not identified because of obfuscation due to bias may cause a decision-making system to not trigger a computer process when it should (e.g., false negative). Additionally or alternatively, semantic similarities that are detected but would not be identified if bias were removed may cause a decision-making system to trigger a computer process when it should not (e.g., false positive). For decision-making systems making a number of decisions to execute or not execute computer processes every second (e.g., fraud monitoring system), the reduction of false negatives and false positives

conserves computer resources (e.g., processing bandwidth, storage space, number of communications, etc.).

[0079] Referring now to FIG. 1, illustrated is a diagram of an example system 100 for debiasing embedding vectors of machine learning models. As shown in FIG. 1, system 100 may include user device 102, modeling system 104, fraud monitoring system 106, decision-making system 108, and communication network 110. Each of the foregoing devices and/or systems may include one or more computing devices configured to communicate (e.g., directly and/or indirectly via communication network 110) with other devices and/or systems in system 100.

[0080] User device 102 may include a computing device configured with a display for visually representing embedding vectors before, during, and/or after debiasing.

[0081] Modeling system 104 may include one or more computing devices (e.g., servers and/or the like) programmed or configured to receive embedding vectors and execute one or more debiasing processes according to methods described herein. In some non-limiting embodiments or aspects, modeling system 104 and user device 102 may be implemented within a single device (e.g., user device 102 may include (e.g., completely, partially, and/or the like) modeling system 104).

[0082] Fraud monitoring system 106 may include one or more computing devices (e.g., servers and/or the like) programmed or configured to receive embedding vectors for use in machine learning models to detect fraud and/or execute one or more fraud-mitigation actions (e.g., freezing accounts, declining transactions, transmitting alerts, etc.).

[0083] Decision-making system 108 may include one or more servers programmed or configured to receive embedding vectors for use in machine learning models to make one or more computer-driven decisions (e.g., execute computer processes, stop executing computer processes, store data, modify data, transmit data, etc.). In some non-limiting embodiments or aspects, decision-making system 108 may include fraud monitoring system 106.

[0084] Communication network 110 may include one or more wired and/or wireless networks. For example, communication network 110 may include a cellular network (e.g., a long-term evolution (LTE®) network, a third generation (3G) network, a fourth generation (4G) network, a fifth generation (5G) network, a code division multiple access (CDMA) network, and/or the like), a public land mobile network (PLMN), a local area network (LAN), a wide area network (WAN), a metropolitan area network (MAN),

a telephone network (e.g., the public switched telephone network (PSTN)), a private network, an ad hoc network, a mesh network, a beacon network, an intranet, the Internet, a fiber optic-based network, a cloud computing network, and/or the like, and/or a combination of these or other types of networks.

[0085] The number and arrangement of devices and networks shown in FIG. 1 are provided as an example. There may be additional devices and/or networks, fewer devices and/or networks, different devices and/or networks, or differently arranged devices and/or networks than those shown in FIG. 1. Furthermore, two or more devices shown in FIG. 1 may be implemented within a single device, or a single device shown in FIG. 1 may be implemented as multiple, distributed devices. Additionally or alternatively, a set of devices (e.g., one or more devices) of system 100 may perform one or more functions described as being performed by another set of devices of system 100.

[0086] Referring now to FIG. 2, illustrated is a diagram of example components of a device 200. Device 200 may correspond to one or more devices of user device 102, modeling system 104, fraud monitoring system 106, decision-making system 108, and/or a communication network 110. In some non-limiting embodiments or aspects, one or more devices of the foregoing may include at least one device 200 and/or at least one component of device 200. As shown in FIG. 2, device 200 may include bus 202, processor 204, memory 206, storage component 208, input component 210, output component 212, and communication interface 214.

[0087] Bus 202 may include a component that permits communication among the components of device 200. In some non-limiting embodiments or aspects, processor 204 may be implemented in hardware, software, or a combination of hardware and software. For example, processor 204 may include a processor (e.g., a central processing unit (CPU), a graphics processing unit (GPU), an accelerated processing unit (APU), etc.), a microprocessor, a digital signal processor (DSP), and/or any processing component (e.g., a field-programmable gate array (FPGA), an application-specific integrated circuit (ASIC), etc.) that can be programmed to perform a function. Memory 206 may include random access memory (RAM), read-only memory (ROM), and/or another type of dynamic or static storage device (e.g., flash memory, magnetic memory, optical memory, etc.) that stores information and/or instructions for use by processor 204.

[0088] Storage component 208 may store information and/or software related to the operation and use of device 200. For example, storage component 208 may include a hard disk (e.g., a magnetic disk, an optical disk, a magneto-optic disk, a solid state disk, etc.), a compact disc (CD), a digital versatile disc (DVD), a floppy disk, a cartridge, a magnetic tape, and/or another type of computer-readable medium, along with a corresponding drive.

[0089] Input component 210 may include a component that permits device 200 to receive information, such as via user input (e.g., a touchscreen display, a keyboard, a keypad, a mouse, a button, a switch, a microphone, a camera, etc.). Additionally or alternatively, input component 210 may include a sensor for sensing information (e.g., a global positioning system (GPS) component, an accelerometer, a gyroscope, an actuator, etc.). Output component 212 may include a component that provides output information from device 200 (e.g., a display, a speaker, one or more light-emitting diodes (LEDs), etc.).

[0090] Communication interface 214 may include a transceiver-like component (e.g., a transceiver, a separate receiver and transmitter, etc.) that enables device 200 to communicate with other devices, such as via a wired connection, a wireless connection, or a combination of wired and wireless connections. Communication interface 214 may permit device 200 to receive information from another device and/or provide information to another device. For example, communication interface 214 may include an Ethernet interface, an optical interface, a coaxial interface, an infrared interface, a radio frequency (RF) interface, a universal serial bus (USB) interface, a Wi-Fi® interface, a cellular network interface, and/or the like.

[0091] Device 200 may perform one or more processes described herein. Device 200 may perform these processes based on processor 204 executing software instructions stored by a computer-readable medium, such as memory 206 and/or storage component 208. A computer-readable medium (e.g., a non-transitory computer-readable medium) is defined herein as a non-transitory memory device. A non-transitory memory device includes memory space located inside of a single physical storage device or memory space spread across multiple physical storage devices.

[0092] Software instructions may be read into memory 206 and/or storage component 208 from another computer-readable medium or from another device via communication interface 214. When executed, software instructions stored in

memory 206 and/or storage component 208 may cause processor 204 to perform one or more processes described herein. Additionally or alternatively, hardwired circuitry may be used in place of or in combination with software instructions to perform one or more processes described herein. Thus, embodiments or aspects described herein are not limited to any specific combination of hardware circuitry and software.

[0093] Memory 206 and/or storage component 208 may include data storage or one or more data structures (e.g., a database, and/or the like). Device 200 may be capable of receiving information from, storing information in, communicating information to, or searching information stored in the data storage or one or more data structures in memory 206 and/or storage component 208. For example, the information may include encryption data, input data, output data, transaction data, account data, or any combination thereof.

[0094] The number and arrangement of components shown in FIG. 2 are provided as an example. In some non-limiting embodiments or aspects, device 200 may include additional components, fewer components, different components, or differently arranged components than those shown in FIG. 2. Additionally or alternatively, a set of components (e.g., one or more components) of device 200 may perform one or more functions described as being performed by another set of components of device 200.

[0095] Referring now to FIG. 3, illustrated is a flow diagram of a method 300 for debiasing embedding vectors of at least one machine learning model (e.g., at least one neural network model and/or the like). One or more steps of method 300 may be executed by one or more processors of modeling system 104. Additionally or alternatively, one or more steps of method 300 may be executed (e.g., completely, partially, and/or the like) by another system, another device, another group of systems, or another group of devices, separate from or including modeling system 104, such as user device 102, fraud monitoring system 106, decision-making system 108, and/or the like. Each step of method 300 may be performed by a same or different processor.

[0096] In step 302, a plurality of embedding vectors may be received. For example, modeling system 104 may receive a plurality of embedding vectors from a neural network model.

[0097] In some non-limiting embodiments or aspects, the embedding vectors may include vector representations of word embeddings, and/or the neural network model may include a natural language processing (NLP) model. For example, a category of

bias for word embeddings may include gender-association of words that are not inherently gendered.

[0098] In some non-limiting embodiments or aspects, the plurality of embedding vectors may be vector representations of merchant identity embeddings generated from customer transaction data. For example, a category of bias for merchant identity embeddings may include location of merchants (e.g., merchants may be associated based on location even if such merchants are not inherently related by type (e.g., merchant category code (MCC) and/or the like)).

[0099] In step 304, at least two clusters may be generated. For example, modeling system 104 may generate two clusters of embedding vectors based on the plurality of embedding vectors.

[0100] In some non-limiting embodiments, the two clusters may include a first cluster of embedding vectors expected to be biased in a first direction (e.g., male-oriented gender bias, first location bias, etc.) and/or a second cluster of embedding vectors expected to be biased in a second direction (e.g., female-oriented gender bias, second location bias, etc.). The direction of bias may be bipolar, and the clusters may be determined based on groups according to one pole of the bias or another. The two clusters may be generated by selecting embedding vectors representing known or statistically biased input variables. Additionally or alternatively, the two clusters may be generated by selecting groups of nearby embedding vectors, as determined visually in an interface or mathematically.

[0101] In step 306, a first mean vector of the first cluster and/or a second mean vector of the second cluster may be determined. For example, modeling system 104 may determine a first mean vector of the first cluster of embedding vectors and a second mean vector of the second cluster of embedding vectors.

[0102] In some non-limiting embodiments or aspects, a two-means approach may be used (e.g., by modeling system 104) to determine the first mean vector and the second mean vector, e.g., a center vector of each cluster may be determined. The two-means method, for any two groups of words, may return the normalized vector of their respective averages. For example, for groups of embedding vectors $s_i \in S$ and $m_i \in M$, modeling system 104 may compute a first mean vector as shown in the following equation:

$$s = \frac{1}{|S|} \sum_i s_i$$

Additionally or alternatively, modeling system 104 may compute a second mean vector as shown in the following equation:

$$m = \frac{1}{|M|} \sum_i m_i$$

[0103] From the above mean vectors, an initial direction may be calculated as shown in the following equation:

$$v = \frac{s - m}{\|s - m\|}$$

[0104] In some non-limiting embodiments or aspects, a user of user device 102 may view a visualization of the embedding vectors (e.g., on the display of user device 102). Additionally or alternatively, the user may provide input to user device 102 to designate another vector within the first cluster as a first mean vector and/or another vector within the second cluster as a second mean vector.

[0105] In some non-limiting embodiments or aspects, modeling system 104 may determine a convex combination of the first cluster as a first mean vector and/or a convex combination of the second cluster as a second mean vector.

[0106] In some non-limiting embodiments or aspects, modeling system 104 may determine an initial (e.g., baseline) bias score (e.g., a value of a bias test statistic) of the plurality of embedding vectors. For example, the bias score may be calculated (e.g., by modeling system 104) from one or more bias test statistics, including, but not limited to, a Word Embedding Association Test (WEAT), an Embedding Coherence Test (ECT), and/or the like. Additionally or alternatively, modeling system 104 may compare the plurality of embedding vectors to embedding vectors representing known biased variables to determine an association (e.g., similarity of bias) between the groups of vectors.

[0107] In some non-limiting embodiments or aspects, modeling system 104 may determine a bias score (e.g., a two-means bias score) based on linearly projecting the embedding vectors of both the first cluster and the second cluster along an initial

direction defined between the first mean vector and the second mean vector. For example, the bias score may be calculated from one or more bias test statistics, including, but not limited to, a WEAT, an ECT, and/or the like. Additionally or alternatively, modeling system 104 may determine a difference between the bias score (e.g., two-means bias score) and the initial bias score.

[0108] In step 308, the first mean vector may be modified by moving the first mean vector toward each vector of the first cluster. For example, modeling system 104 may modify the first mean vector by moving the first mean vector toward each embedding vector of the first cluster to provide a plurality of first candidate vectors, and by linearly projecting the embedding vectors of both the first cluster and the second cluster along a first direction defined between the second mean vector and each first candidate vector to determine a bias score associated with each first candidate vector.

[0109] In some non-limiting embodiments or aspects, moving the first mean vector toward each respective vector in the first cluster may comprise identifying a convex combination of the first mean vector and the respective vector resulting in an improvement in the bias score. Additionally or alternatively, the moving of the first mean vector may include using Golden Section Search (GSS) to optimize an improvement in the bias score by evaluating candidate vectors (e.g., convex combinations of the first mean vector and respective embedding vectors of the first cluster).

[0110] In step 310, the first mean vector may be replaced with a first candidate vector based on the bias score of the first candidate vector. For example, modeling system 104 may replace the first mean vector with a first candidate vector of the plurality of first candidate vectors based on the bias score of the first candidate vector.

[0111] In some non-limiting embodiments or aspects, modeling system 104 may prioritize identifying a first candidate vector having a local extremum of value of the bias score (e.g., a lower WEAT score, a higher ECT score, etc.). Additionally or alternatively, modeling system 104 may replace the first mean vector with the first candidate vector of the plurality of first candidate vectors based on maximizing a difference between the initial bias score and the bias score of the first candidate vector (e.g., a greatest decrease in WEAT score, a greatest increase in ECT score, etc.). The chosen first candidate vector may become the new first mean vector.

[0112] In step 312, the second mean vector may be modified by moving the second mean vector toward each vector of the second cluster. For example, modeling

system 104 may modify the second mean vector by moving the second mean vector toward each embedding vector of the second cluster to provide a plurality of second candidate vectors, and by linearly projecting the embedding vectors of both the first cluster and the second cluster along a second direction defined between the first mean vector and each second candidate vector to determine a bias score associated with each second candidate vector.

[0113] In some non-limiting embodiments or aspects, moving the second mean vector toward each respective vector in the second cluster may comprise identifying a convex combination of the second mean vector and the respective vector resulting in an improvement in the bias score. Additionally or alternatively, the moving of the second mean vector may include using GSS to optimize an improvement in the bias score by evaluating candidate vectors (e.g., convex combinations of the second mean vector and respective embedding vectors of the second cluster).

[0114] In step 314, the second mean vector may be replaced with a second candidate vector based on the bias score of the second candidate vector. For example, modeling system 104 may replace the second mean vector with a second candidate vector of the plurality of second candidate vectors based on the bias score of the second candidate vector.

[0115] In some non-limiting embodiments or aspects, modeling system 104 may prioritize identifying a second candidate vector having a local extremum of value of the bias score (e.g., a lower WEAT score, a higher ECT score, etc.). Additionally or alternatively, modeling system 104 may replace the second mean vector with the second candidate vector of the plurality of second candidate vectors based on maximizing a difference between the initial bias score and the bias score of the second candidate vector (e.g., a greatest decrease in WEAT score, a greatest increase in ECT score, etc.).

[0116] In some non-limiting embodiments, steps 308, 310, 312, and 314 may be repeated until an extremum of the bias score is reached (e.g., convergence on a maximum or minimum value of a bias test statistic). For example, modeling system 104 may repeat the modifying of the first mean vector in step 308, the replacing of the first mean vector in step 310, the modifying of the second mean vector in step 312, and the replacing of the second mean vector in step 314 until an extremum of the bias score is reached (e.g., a maximum ECT score, a minimum WEAT score, etc.).

[0117] In step 316, the plurality of embedding vectors may be linearly projected along a direction defined between the first mean vector and the second mean vector (e.g., after being moved as described above). For example, modeling system 104 may, in response to reaching the extremum of the bias score, linearly project each embedding vector of the plurality of embedding vectors along a direction defined between the first mean vector and the second mean vector to debias the plurality of embedding vectors.

[0118] In some non-limiting embodiments or aspects, modeling system 104 may communicate the (debiased) embedding vectors to at least one of fraud monitoring system 106 and/or decision-making system 108.

[0119] In some non-limiting embodiments or aspects, decision-making system 108 may use the debiased embedding vectors to modify (e.g., train, retrain, and/or the like) a machine learning model used to execute one or more computer-driven processes.

[0120] In some non-limiting embodiments or aspects, for embedding vectors including vector representations of word embeddings generated from document data, a system for natural language processing may modify (e.g., train, retrain, and/or the like) a machine learning model used for text-based search queries, machine translation, and/or the like. Use of the debiased word embeddings may improve the accuracy and efficiency of said search queries and machine translation.

[0121] In some non-limiting embodiments or aspects, for embedding vectors including vector representations of merchant identity embeddings generated from customer transaction data, fraud monitoring system 106 may modify (e.g., train, retrain, and/or the like) a machine learning model of fraud monitoring system 106 using the plurality of embedding vectors that have been debiased, so as to further increase the accuracy and efficiency of fraud detection and mitigation.

[0122] In some non-limiting embodiments or aspects, for embedding vectors including vector representations of merchant identity embeddings generated from customer transaction data, a recommendation system may modify (e.g., train, retrain, and/or the like) a machine learning model used for making merchant recommendations to users (e.g., in a mobile application, a web browser, in push communications, etc.). Use of the debiased merchant identity embeddings may improve the accuracy and efficiency of said merchant recommendations.

[0123] Referring now to FIGS. 4A-4D, illustrated are diagrams of an exemplary implementation of the method shown in FIG. 3. One or more steps of an

implementation 400 may be executed by one or more processors of modeling system 104. Additionally or alternatively, one or more steps of implementation 400 may be executed (e.g., completely, partially, and/or the like) by another system, another device, another group of systems, or another group of devices, separate from or including modeling system 104, such as user device 102, fraud monitoring system 106, decision-making system 108, and/or the like. Each step of implementation 400 may be performed by a same or different processor.

[0124] As shown in FIG. 4A, in some non-limiting embodiments or aspects, a plurality of embedding vectors 401, 403 may be received. (For clarity, each embedding vector 401, 403 is illustrated as a dot rather than a ray. Each dot may represent the endpoint of the respective vector.) For example, modeling system 104 may receive a plurality of embedding vectors 401, 403, as described herein. Embedding vectors 401, 403 may include first embedding vectors 401 in a first category and second embedding vectors 403 in a second category, as described herein. In some non-limiting embodiments or aspects, the first category and the second category each may be a respective subcategory of a category of bias. For example, the first category may include a first location and the second category may include a second location, and location may be a category of bias for merchant embedding vectors. For example, the first category may include a first gender (e.g., male) and the second category may include a second gender (e.g., female), and gender may be a category of bias word embedding vectors.

[0125] In some non-limiting embodiments or aspects, at least two clusters may be generated. For example, modeling system 104 may generate two clusters 421, 423 based on first embedding 401 and second embedding vector 403, as described herein.

[0126] In some non-limiting embodiments or aspects, first mean vector 411 of first cluster 421 and/or second mean vector 413 of second cluster 423 may be determined. For example, modeling system 104 may determine first mean vector 411 of first cluster 421 of first embedding vectors 401 and second mean vector 413 of second cluster 423 of second embedding vectors 403, as described herein. In some non-limiting embodiments or aspects, an initial direction 431a may be determined based on first mean vector 411 and second mean vector 413, as described herein.

[0127] In some non-limiting embodiments or aspects, second mean vector 413 may be modified by moving second mean vector 413 toward each (second) embedding

vector 403 of second cluster 423, as described herein. For example, modeling system 104 may modify second mean vector 413, as described herein.

[0128] As shown in FIG. 4B, in some non-limiting embodiments or aspects, second mean vector 413 may be replaced with (second) candidate vector 413a based on the bias score of (second) candidate vector 413a, as described herein. For example, modeling system 104 may replace second mean vector 413 with (second) candidate vector 413a of a plurality of (second) candidate vectors based on the bias score of (second) candidate vector 413a, as described herein. In some non-limiting embodiments or aspects, modified direction 431b may be determined based on first mean vector 411 and (second) candidate vector 413a, as described herein.

[0129] As shown in FIG. 4C, in some non-limiting embodiments or aspects, first mean vector 411 may be modified by moving first mean vector 411 toward each (first) embedding vector 401 of first cluster 421, as described herein. For example, modeling system 104 may modify first mean vector 411, as described herein.

[0130] As shown in FIG. 4D, in some non-limiting embodiments or aspects, first mean vector 411 may be replaced with (first) candidate vector 411a based on the bias score of (first) candidate vector 411a, as described herein. For example, modeling system 104 may replace first mean vector 411 with (first) candidate vector 411a of a plurality of (first) candidate vectors based on the bias score of (first) candidate vector 411a, as described herein. In some non-limiting embodiments or aspects, a further modified direction 431c may be determined based on first candidate vector 411a and second candidate vector 413a, as described herein.

[0131] In some non-limiting embodiments or aspects, the aforementioned process may be repeated until an extremum of the bias score is reached (e.g., convergence on a maximum or minimum value of a bias test statistic), as described herein.

[0132] Referring now to FIG. 5, illustrated is a diagram of an exemplary implementation 500 of the method shown in FIG. 3. As shown in FIG. 5, implementation 500 includes customer data 501, first merchant data 503, second merchant data 505, embedding datasets 511, other transaction data 513, customer embedding vector 521, first merchant embedding vector 523, second merchant embedding vector 525, other data 527, machine learning model 530, first prediction 531, and second prediction 533. In some non-limiting embodiments or aspects, implementation 500 may be implemented (e.g., completely, partially, and/or the like) by modeling system 104. Additionally or alternatively, implementation 500 may be

implemented (e.g., completely, partially, and/or the like) by another system, another device, another group of systems, or another group of devices, separate from or including modeling system 104, such as user device 102, fraud monitoring system 106, decision-making system 108, and/or the like. For example, as shown in FIG. 5 for the purpose of illustration and not limitation, implementation 500 may be implemented partially by modeling system 104 and partially by fraud monitoring system 106 and/or decision-making system 108.

[0133] In some non-limiting embodiments or aspects, customer data 501 may include transaction data (e.g., payment transaction data associated with a plurality of payment transactions) associated with a customer (e.g., a payment account of the customer). Additionally or alternatively, first merchant data 503 may include transaction data (e.g., payment transaction data associated with a plurality of payment transactions) associated with a first merchant (e.g., a first restaurant and/or the like). Additionally or alternatively, second merchant data 505 may include transaction data (e.g., payment transaction data associated with a plurality of payment transactions) associated with a second merchant (e.g., a second restaurant and/or the like).

[0134] In some non-limiting embodiments or aspects, embedding datasets 511 may include data based on customer data 501 (or a portion thereof), first merchant data 503 (or a portion thereof), and/or second merchant data 505 (or a portion thereof). Additionally or alternatively, other transaction data 513 may include data based on customer data 501 (or a portion thereof), first merchant data 503 (or a portion thereof), and/or second merchant data 505 (or a portion thereof).

[0135] In some non-limiting embodiments or aspects, customer embedding vector 521, first merchant embedding vector 523, and/or second merchant embedding vector 525 may be generated, e.g., by a neural network model and/or NLP model, as described herein. Additionally or alternatively, at least one of customer embedding vector 521, first merchant embedding vector 523, and/or second merchant embedding vector 525 may be modified (e.g., linearly projected) to remove bias, as described herein.

[0136] In some non-limiting embodiments or aspects, customer embedding vector 521, first merchant embedding vector 523, and/or second merchant embedding vector 525 (e.g., after being modified/linearly projected to remove bias, as described herein) may be communicated to and/or inputted to machine learning model 530, as described herein. For example, machine learning model 530 may include a neural network, a

deep neural network (DNN) a fraud detection model, a recommendation model (e.g., merchant recommendation model, restaurant recommendation model, product recommendation model, and/or the like), and/or the like.

[0137] In some non-limiting embodiments or aspects, customer embedding vector 521, first merchant embedding vector 523, and/or second merchant embedding vector 525 may be combined with (e.g., appended to and/or concatenated with) other data 527 based on other transaction data 513 before or after being communicated to machine learning model 530 (e.g., before being inputted to machine learning model 530).

[0138] In some non-limiting embodiments or aspects, machine learning model 530 may be modified (e.g., trained, retrained, and/or the like) based on the embedding vectors (e.g., customer embedding vector 521, first merchant embedding vector 523, and/or second merchant embedding vector 525, each of which may have been modified/linearly projected to remove bias, as described herein). Additionally or alternatively, machine learning model 530 (e.g., after having been trained) may generate at least one prediction based on the embedding vectors (e.g., customer embedding vector 521, first merchant embedding vector 523, and/or second merchant embedding vector 525, each of which may have been modified/linearly projected to remove bias, as described herein). For example, machine learning model 530 may generate first prediction 531 indicating that the customer would likely prefer the first merchant (e.g., first restaurant) and/or second prediction 533 indicating that the customer would likely prefer the second merchant (e.g., second restaurant) based on the embedding vectors.

[0139] FIGS. 6A-6C are graphical visualizations of exemplary vectors of an exemplary implementation of the process shown in FIG. 3. In each visualization, embedding vectors that have a relatively high number of dimensions (e.g., greater than two dimensions, such as 50 dimensions or more) are projected into two-dimensional space to enable visualizing the embedding vectors. For clarity, each embedding vector is illustrated as a point rather than a ray. Each point may represent the endpoint of the respective vector. Each point is labeled with a location (e.g., "LA" for Los Angeles or "BAY" for Bay Area) and the name of a merchant (e.g., restaurant) represented by the respective embedding vector. Circular points are embedding vectors used for training, and square points are embedding vectors used for

evaluation. The direction of bias (e.g., with location as the category of bias) is indicated by the line (labeled "Location") from the origin (e.g., "Origin").

[0140] As shown in FIG. 6A, the embedding vectors may be projected into two-dimensional space such that the horizontal and vertical axes of the two-dimensional space are the dimensions with highest variance (e.g., as determined by principal component analysis (PCA)). For the purpose of illustration and not limitation, in this two-dimensional space, the direction of (location) bias from the origin is towards the top and slightly to the left within the visualization. For the purpose of illustration and not limitation, the embedding vectors for merchants in Los Angeles are generally clustered in the bottom portion of the visualization, and the embedding vectors for merchants in the Bay Area are generally clustered in the top portion of the visualization. In some non-limiting embodiments or aspects, each merchant may be a restaurant with a cuisine of one of seven types: Chinese, Ramen, Korean BBQ, Diner, Donuts, Pizza, and Burgers/Hotdogs. However, as shown in FIG. 6A, restaurants having cuisine of the same type are not initially clustered together.

[0141] As shown in FIG. 6B, the two-dimensional visualization space may be rotated so that the direction of bias is aligned with the horizontal axis. Additionally or alternatively, the dimension of the (higher-dimensional) embedding vector space with highest variance (e.g., as determined by PCA) may be aligned with the vertical axis. For the purpose of illustration and not limitation, the direction of (location) bias from the origin is straight to the right of the visualization. For the purpose of illustration and not limitation, the embedding vectors for merchants in Los Angeles are generally clustered in the left portion of the visualization, and the embedding vectors for merchants in the Bay Area are generally clustered in the right portion of the visualization. However, as shown in FIG. 6B, restaurants having cuisine of the same type are not initially clustered together.

[0142] As shown in FIG. 6C, the embedding vectors are modified/linearly projected to remove bias, as described herein. For the purpose of illustration and not limitation, the (location) bias has been removed (e.g., the label Location now overlaps with the label Origin because there is not direction of bias). Additionally, as shown in FIG. 6C, restaurants having cuisine of the same type are now clustered together. For example, cluster 601 includes two Donuts restaurants, even though they are in different locations (Los Angeles and the Bay Area). Cluster 603 includes two Burgers/Hotdogs restaurants, even though they are in different locations. Cluster 605 includes two

Pizza restaurants, even though they are in different locations. Cluster 607 includes two Ramen restaurants, even though they are in different locations. Cluster 609 includes two Diner restaurants, even though they are in different locations. Cluster 611 includes two Chinese restaurants, even though they are in different locations. Cluster 613 includes two Korean BBQ restaurants, even though they are in different locations.

[0143] For the purpose of illustration and not limitation, Table 1 shows three bias scores (WEAT, ECT, and Natural Language Inference (NLI)) for the disclosed methodology (“Iterative Subspace”) compared to embeddings without any correction for bias (“Baseline”) and three different methodologies for removing bias: PCA (“PCA”), two-means (“2-means”), and classification normal (“Classification”) for the task of removing gender bias from word embeddings. For ECT and NLI, higher values may be desirable (e.g., less bias), and for WEAT, lower values may be desirable (e.g., less bias). As can be seen from Table 1, the disclosed methodology (“Iterative Subspace”) outperforms the other methodologies for removing bias.

Method	ECT (occ)	WEAT (adj)	NLI Test
Baseline	0.773	1.587	0.297
PCA	0.905	1.17	0.346
2-means	0.912	1.102	0.379
Classification (1 step)	0.872	0.951	0.383
Iterative Subspace	0.966	0.902	0.386

Table 1

[0144] Referring now to FIG. 7, FIG. 7 is a diagram of an exemplary environment 700 in which systems, products, and/or methods, as described herein, may be implemented, according to non-limiting embodiments or aspects of the presently disclosed subject matter. As shown in FIG. 7, environment 700 includes transaction service provider system 702, issuer system 704, customer device 706, merchant system 708, acquirer system 710, and communication network 712. In some non-limiting embodiments or aspects, at least one of (e.g., all of) user device 102, modeling system 104, fraud monitoring system 106, and/or decision-making system 108 may be the same as, similar to, and/or part of transaction service provider system 702. In some non-limiting embodiments or aspects, at least one of user device 102, modeling system 104, fraud monitoring system 106, and/or decision-making system 108 may

be the same as, similar to, and/or part of another system, another device, another group of systems, or another group of devices, separate from or including transaction service provider system 702, such as issuer system 704, customer device 706, merchant system 708, acquirer system 710, and/or the like. For example, transaction service provider system 702 may include each of modeling system 104, fraud monitoring system 106, and decision-making system 108, and at least one of issuer system 704, customer device 706, merchant system 708, and/or acquirer system 710 may include user device 102. Additionally or alternatively, transaction service provider system 702 may include modeling system 104, and at least one of issuer system 704, customer device 706, merchant system 708, and/or acquirer system 710 may include one or more of user device 102, fraud monitoring system 106, and/or decision-making system 108. Additionally or alternatively, transaction service provider system 702 may include modeling system 104 and fraud monitoring system 106, and at least one of issuer system 704, customer device 706, merchant system 708, and/or acquirer system 710 may include one or more of user device 102 and/or decision-making system 108.

[0145] Transaction service provider system 702 may include one or more devices capable of receiving information from and/or communicating information to issuer system 704, customer device 706, merchant system 708, and/or acquirer system 710 via communication network 712. For example, transaction service provider system 702 may include a computing device, such as a server (e.g., a transaction processing server), a group of servers, and/or other like devices. In some non-limiting embodiments or aspects, transaction service provider system 702 may be associated with a transaction service provider as described herein. In some non-limiting embodiments or aspects, transaction service provider system 702 may be in communication with a data storage device, which may be local or remote to transaction service provider system 702. In some non-limiting embodiments or aspects, transaction service provider system 702 may be capable of receiving information from, storing information in, communicating information to, or searching information stored in the data storage device.

[0146] Issuer system 704 may include one or more devices capable of receiving information and/or communicating information to transaction service provider system 702, customer device 706, merchant system 708, and/or acquirer system 710 via communication network 712. For example, issuer system 704 may include a

computing device, such as a server, a group of servers, and/or other like devices. In some non-limiting embodiments or aspects, issuer system 704 may be associated with an issuer institution as described herein. For example, issuer system 704 may be associated with an issuer institution that issued a credit account, debit account, credit card, debit card, and/or the like to a user associated with customer device 706.

[0147] Customer device 706 may include one or more devices capable of receiving information from and/or communicating information to transaction service provider system 702, issuer system 704, merchant system 708, and/or acquirer system 710 via communication network 712. Additionally or alternatively, each customer device 706 may include a device capable of receiving information from and/or communicating information to other user devices 706 via communication network 712, another network (e.g., an ad hoc network, a local network, a private network, a virtual private network, and/or the like), and/or any other suitable communication technique. For example, customer device 706 may include a client device and/or the like. In some non-limiting embodiments or aspects, customer device 706 may or may not be capable of receiving information (e.g., from merchant system 708 or from another customer device 706) via a short-range wireless communication connection (e.g., an NFC communication connection, an RFID communication connection, a Bluetooth® communication connection, a Zigbee® communication connection, and/or the like), and/or communicating information (e.g., to merchant system 708) via a short-range wireless communication connection.

[0148] Merchant system 708 may include one or more devices capable of receiving information from and/or communicating information to transaction service provider system 702, issuer system 704, customer device 706, and/or acquirer system 710 via communication network 712. Merchant system 708 may also include a device capable of receiving information from customer device 706 via communication network 712, a communication connection (e.g., an NFC communication connection, an RFID communication connection, a Bluetooth® communication connection, a Zigbee® communication connection, and/or the like) with customer device 706, and/or the like, and/or communicating information to customer device 706 via communication network 712, the communication connection, and/or the like. In some non-limiting embodiments or aspects, merchant system 708 may include a computing device, such as a server, a group of servers, a client device, a group of client devices, and/or other like devices. In some non-limiting embodiments or aspects, merchant system 708

may be associated with a merchant, as described herein. In some non-limiting embodiments or aspects, merchant system 708 may include one or more client devices. For example, merchant system 708 may include a client device that allows a merchant to communicate information to transaction service provider system 702. In some non-limiting embodiments or aspects, merchant system 708 may include one or more devices, such as computers, computer systems, and/or peripheral devices capable of being used by a merchant to conduct a transaction with a user. For example, merchant system 708 may include a POS device and/or a POS system.

[0149] Acquirer system 710 may include one or more devices capable of receiving information from and/or communicating information to transaction service provider system 702, issuer system 704, customer device 706, and/or merchant system 708 via communication network 712. For example, acquirer system 710 may include a computing device, a server, a group of servers, and/or the like. In some non-limiting embodiments or aspects, acquirer system 710 may be associated with an acquirer as described herein.

[0150] Communication network 712 may include one or more wired and/or wireless networks. For example, communication network 712 may include a cellular network (e.g., an LTE® network, a 3G network, a 4G network, a 5G network, a CDMA network, and/or the like), a PLMN, LAN, a WAN, a MAN, a telephone network (e.g., the PSTN), a private network (e.g., a private network associated with a transaction service provider), an ad hoc network, an intranet, the Internet, a fiber optic-based network, a cloud computing network, and/or the like, and/or a combination of these or other types of networks.

[0151] In some non-limiting embodiments or aspects, processing a transaction may include generating and/or communicating at least one transaction message (e.g., authorization request, authorization response, any combination thereof, and/or the like). For example, a client device (e.g., customer device 706, a POS device of merchant system 708, and/or the like) may initiate the transaction, e.g., by generating an authorization request. Additionally or alternatively, the client device (e.g., customer device 706, at least one device of merchant system 708, and/or the like) may communicate the authorization request. For example, customer device 706 may communicate the authorization request to merchant system 708 and/or a payment gateway (e.g., a payment gateway of transaction service provider system 702, a third-party payment gateway separate from transaction service provider system 702, and/or

the like). Additionally or alternatively, merchant system 708 (e.g., a POS device thereof) may communicate the authorization request to acquirer system 710 and/or a payment gateway. In some non-limiting embodiments or aspects, acquirer system 710 and/or a payment gateway may communicate the authorization request to transaction service provider system 702 and/or issuer system 704. Additionally or alternatively, transaction service provider system 702 may communicate the authorization request to issuer system 704. In some non-limiting embodiments or aspects, issuer system 704 may determine an authorization decision (e.g., authorize, decline, and/or the like) based on the authorization request. For example, the authorization request may cause issuer system 704 to determine the authorization decision based thereof. In some non-limiting embodiments or aspects, issuer system 704 may generate an authorization response based on the authorization decision. Additionally or alternatively, issuer system 704 may communicate the authorization response. For example, issuer system 704 may communicate the authorization response to transaction service provider system 702 and/or a payment gateway. Additionally or alternatively, transaction service provider system 702 and/or a payment gateway may communicate the authorization response to acquirer system 710, merchant system 708, and/or customer device 706. Additionally or alternatively, acquirer system 710 may communicate the authorization response to merchant system 708 and/or a payment gateway. Additionally or alternatively, a payment gateway may communicate the authorization response to merchant system 708 and/or customer device 706. Additionally or alternatively, merchant system 708 may communicate the authorization response to customer device 706. In some non-limiting embodiments or aspects, merchant system 708 may receive (e.g., from acquirer system 710 and/or a payment gateway) the authorization response. Additionally or alternatively, merchant system 708 may complete the transaction based on the authorization response (e.g., provide, ship, and/or deliver goods and/or services associated with the transaction; fulfill an order associated with the transaction; any combination thereof; and/or the like).

[0152] For the purpose of illustration, processing a transaction may include generating a transaction message (e.g., authorization request and/or the like) based on an account identifier of a customer (e.g., associated with customer device 706 and/or the like) and/or transaction data associated with the transaction. For example, merchant system 708 (e.g., a client device of merchant system 708, a POS device of

merchant system 708, and/or the like) may initiate the transaction, e.g., by generating an authorization request (e.g., in response to receiving the account identifier from a portable financial device of the customer and/or the like). Additionally or alternatively, merchant system 708 may communicate the authorization request to acquirer system 710. Additionally or alternatively, acquirer system 710 may communicate the authorization request to transaction service provider system 702. Additionally or alternatively, transaction service provider system 702 may communicate the authorization request to issuer system 704. Issuer system 704 may determine an authorization decision (e.g., authorize, decline, and/or the like) based on the authorization request, and/or issuer system 704 may generate an authorization response based on the authorization decision and/or the authorization request. Additionally or alternatively, issuer system 704 may communicate the authorization response to transaction service provider system 702. Additionally or alternatively, transaction service provider system 702 may communicate the authorization response to acquirer system 710, which may communicate the authorization response to merchant system 708.

[0153] For the purpose of illustration, clearing and/or settlement of a transaction may include generating a message (e.g., clearing message, settlement message, and/or the like) based on an account identifier of a customer (e.g., associated with customer device 706 and/or the like) and/or transaction data associated with the transaction. For example, merchant system 708 may generate at least one clearing message (e.g., a plurality of clearing messages, a batch of clearing messages, and/or the like). Additionally or alternatively, merchant system 708 may communicate the clearing message(s) to acquirer system 710. Additionally or alternatively, acquirer system 710 may communicate the clearing message(s) to transaction service provider system 702. Additionally or alternatively, transaction service provider system 702 may communicate the clearing message(s) to issuer system 704. Additionally or alternatively, issuer system 704 may generate at least one settlement message based on the clearing message(s). Additionally or alternatively, issuer system 704 may communicate the settlement message(s) and/or funds to transaction service provider system 702 (and/or a settlement bank system associated with transaction service provider system 702). Additionally or alternatively, transaction service provider system 702 (and/or the settlement bank system) may communicate the settlement message(s) and/or funds to acquirer system 710, which may communicate the

settlement message(s) and/or funds to merchant system 708 (and/or an account associated with merchant system 708).

[0154] The number and arrangement of systems, devices, and/or networks shown in FIG. 7 are provided as an example. There may be additional systems, devices, and/or networks; fewer systems, devices, and/or networks; different systems, devices, and/or networks; and/or differently arranged systems, devices, and/or networks than those shown in FIG. 7. Furthermore, two or more systems or devices shown in FIG. 7 may be implemented within a single system or device, or a single system or device shown in FIG. 7 may be implemented as multiple, distributed systems or devices. Additionally or alternatively, a set of systems (e.g., one or more systems) or a set of devices (e.g., one or more devices) of environment 700 may perform one or more functions described as being performed by another set of systems or another set of devices of environment 700.

[0155] Although the disclosure has been described in detail for the purpose of illustration based on what is currently considered to be the most practical and non-limiting embodiments or aspects, it is to be understood that such detail is solely for that purpose and that the disclosure is not limited to the disclosed embodiments or aspects, but, on the contrary, is intended to cover modifications and equivalent arrangements that are within the spirit and scope of the appended claims. For example, it is to be understood that the present disclosure contemplates that, to the extent possible, one or more features of any embodiment or aspect can be combined with one or more features of any other embodiment or aspect.

WHAT IS CLAIMED IS

1. A computer-implemented method comprising:
 - receiving, with at least one processor, a plurality of embedding vectors from a neural network model;
 - generating, with the at least one processor, two clusters of embedding vectors based on the plurality of embedding vectors, the two clusters comprising a first cluster of embedding vectors expected to be biased in a first direction and a second cluster expected to be biased in a second direction;
 - determining, with the at least one processor, a first mean vector of the first cluster of embedding vectors and a second mean vector of the second cluster of embedding vectors;
 - modifying, with the at least one processor, the first mean vector by moving the first mean vector toward each embedding vector of the first cluster of embedding vectors to provide a plurality of first candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a first direction defined between the second mean vector and each first candidate vector to determine a bias score associated with each first candidate vector;
 - replacing, with the at least one processor, the first mean vector with a first candidate vector of the plurality of first candidate vectors based on the bias score of the first candidate vector;
 - modifying, with the at least one processor, the second mean vector by moving the second mean vector toward each embedding vector of the second cluster of embedding vectors to provide a plurality of second candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a second direction defined between the first mean vector and each second candidate vector to determine a bias score associated with each second candidate vector;
 - replacing, with the at least one processor, the second mean vector with a second candidate vector of the plurality of second candidate vectors based on the bias score of the second candidate vector;
 - repeating, with the at least one processor, the modifying of the first mean vector, the replacing of the first mean vector, the modifying of the second mean vector, and the replacing of the second mean vector until an extremum of the bias score is reached; and

in response to reaching the extremum of the bias score, linearly projecting, with the at least one processor, each embedding vector of the plurality of embedding vectors along a direction defined between the first mean vector and the second mean vector to debias the plurality of embedding vectors.

2. The computer-implemented method of claim 1, wherein the plurality of embedding vectors are vector representations of word embeddings and the neural network model comprises a natural language processing (NLP) model.

3. The computer-implemented method of claim 1, wherein the plurality of embedding vectors are vector representations of merchant identity embeddings generated from customer transaction data.

4. The computer-implemented method of claim 3, further comprising modifying, with the at least one processor, a machine learning model of a fraud monitoring system using the plurality of embedding vectors that have been debiased by linearly projecting each embedding vector of the plurality of embedding vectors using the direction defined between the first mean vector and the second mean vector.

5. The computer-implemented method of claim 1, wherein the bias score is calculated from a Word Embedding Association Test (WEAT).

6. The computer-implemented method of claim 1, wherein the bias score is calculated from an Embedding Coherence Test (ECT).

7. The computer-implemented method of claim 1, further comprising, before modifying the first mean vector and the second mean vector, determining, with the at least one processor, an initial bias score by linearly projecting the embedding vectors of both the first cluster and the second cluster along an initial direction defined between the first mean vector and the second mean vector, wherein:

the replacing of the first mean vector with the first candidate vector of the plurality of first candidate vectors is based on maximizing a difference between the initial bias score and the bias score of the first candidate vector; and

the replacing of the second mean vector with the second candidate vector of the plurality of second candidate vectors is based on maximizing a difference between the initial bias score and the bias score of the second candidate vector.

8. A system comprising a server comprising at least one processor, the server programmed or configured to:

receive a plurality of embedding vectors from a neural network model;

generate two clusters of embedding vectors based on the plurality of embedding vectors, the two clusters comprising a first cluster of embedding vectors expected to be biased in a first direction and a second cluster expected to be biased in a second direction;

determine a first mean vector of the first cluster of embedding vectors and a second mean vector of the second cluster of embedding vectors;

modify the first mean vector by moving the first mean vector toward each embedding vector of the first cluster of embedding vectors to provide a plurality of first candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a first direction defined between the second mean vector and each first candidate vector to determine a bias score associated with each first candidate vector;

replace the first mean vector with a first candidate vector of the plurality of first candidate vectors based on the bias score of the first candidate vector;

modify the second mean vector by moving the second mean vector toward each embedding vector of the second cluster of embedding vectors to provide a plurality of second candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a second direction defined between the first mean vector and each second candidate vector to determine a bias score associated with each second candidate vector;

replace the second mean vector with a second candidate vector of the plurality of second candidate vectors based on the bias score of the second candidate vector;

repeat the modifying of the first mean vector, the replacing of the first mean vector, the modifying of the second mean vector, and the replacing of the second mean vector until an extremum of the bias score is reached; and

in response to reaching the extremum of the bias score, linearly project each embedding vector of the plurality of embedding vectors along a direction defined between the first mean vector and the second mean vector to debias the plurality of embedding vectors.

9. The system of claim 8, wherein the plurality of embedding vectors are vector representations of word embeddings and the neural network model comprises a natural language processing (NLP) model.

10. The system of claim 8, wherein the plurality of embedding vectors are vector representations of merchant identity embeddings generated from customer transaction data.

11. The system of claim 10, wherein the server is further programmed or configured to modify a machine learning model of a fraud monitoring system using the plurality of embedding vectors that have been debiased by linearly projecting each embedding vector of the plurality of embedding vectors using the direction defined between the first mean vector and the second mean vector.

12. The system of claim 8, wherein the bias score is calculated from a Word Embedding Association Test (WEAT).

13. The system of claim 8, wherein the bias score is calculated from an Embedding Coherence Test (ECT).

14. The system of claim 8, wherein the server is further programmed or configured to, before modifying the first mean vector and the second mean vector, determine an initial bias score by linearly projecting the embedding vectors of both the first cluster and the second cluster along an initial direction defined between the first mean vector and the second mean vector, and wherein:

the replacing of the first mean vector with the first candidate vector of the plurality of first candidate vectors is based on maximizing a difference between the initial bias score and the bias score of the first candidate vector; and

the replacing of the second mean vector with the second candidate vector of the plurality of second candidate vectors is based on maximizing a difference between the initial bias score and the bias score of the second candidate vector.

15. A computer program product comprising at least one non-transitory computer-readable medium including program instructions that, when executed by at least one processor, cause the at least one processor to:

receive a plurality of embedding vectors from a neural network model;

generate two clusters of embedding vectors based on the plurality of embedding vectors, the two clusters comprising a first cluster of embedding vectors expected to be biased in a first direction and a second cluster expected to be biased in a second direction;

determine a first mean vector of the first cluster of embedding vectors and a second mean vector of the second cluster of embedding vectors;

modify the first mean vector by moving the first mean vector toward each embedding vector of the first cluster of embedding vectors to provide a plurality of first candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a first direction defined between the second mean vector and each first candidate vector to determine a bias score associated with each first candidate vector;

replace the first mean vector with a first candidate vector of the plurality of first candidate vectors based on the bias score of the first candidate vector;

modify the second mean vector by moving the second mean vector toward each embedding vector of the second cluster of embedding vectors to provide a plurality of second candidate vectors and linearly projecting the embedding vectors of both the first cluster and the second cluster along a second direction defined between the first mean vector and each second candidate vector to determine a bias score associated with each second candidate vector;

replace the second mean vector with a second candidate vector of the plurality of second candidate vectors based on the bias score of the second candidate vector;

repeat the modifying of the first mean vector, the replacing of the first mean vector, the modifying of the second mean vector, and the replacing of the second mean vector until an extremum of the bias score is reached; and

in response to reaching the extremum of the bias score, linearly project each embedding vector of the plurality of embedding vectors along a direction defined between the first mean vector and the second mean vector to debias the plurality of embedding vectors.

16. The computer program product of claim 15, wherein the plurality of embedding vectors are vector representations of word embeddings and the neural network model comprises a natural language processing (NLP) model.

17. The computer program product of claim 15, wherein the plurality of embedding vectors are vector representations of merchant identity embeddings generated from customer transaction data, and

wherein the instructions, when executed by at least one processor, further cause the at least one processor to modify a machine learning model of a fraud monitoring system using the plurality of embedding vectors that have been debiased by linearly projecting each embedding vector of the plurality of embedding vectors using the direction defined between the first mean vector and the second mean vector.

18. The computer program product of claim 15, wherein the bias score is calculated from a Word Embedding Association Test (WEAT).

19. The computer program product of claim 15, wherein the bias score is calculated from an Embedding Coherence Test (ECT).

20. The computer program product of claim 15, wherein the instructions, when executed by at least one processor, further cause the at least one processor to, before modifying the first mean vector and the second mean vector, determine an initial bias score by linearly projecting the embedding vectors of both the first cluster and the second cluster along an initial direction defined between the first mean vector and the second mean vector, and wherein:

the replacing of the first mean vector with the first candidate vector of the plurality of first candidate vectors is based on maximizing a difference between the initial bias score and the bias score of the first candidate vector; and

the replacing of the second mean vector with the second candidate vector of the plurality of second candidate vectors is based on maximizing a difference between the initial bias score and the bias score of the second candidate vector.