

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;
 - determining, with the at least one processor, a bias vector by:
 - 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 the 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; and

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,

wherein the bias vector is based on a direction defined between the first mean vector and the second mean vector;

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 the bias vector to debias the plurality of embedding vectors to provide a plurality of debiased embedding vectors;

training, with at least one processor, a machine learning model of a fraud monitoring system based on the plurality of debiased embedding vectors;

receiving, with at least one processor, an authorization request associated with a transaction; and

declining, with at least one processor, the transaction associated with the authorization request based on the machine learning model of the fraud monitoring system as trained based on the plurality of debiased embedding vectors.

2. 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.

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

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

5. 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.

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

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

determine a bias vector by:

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

modifying 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 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 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 the bias score associated with each second candidate vector;

replacing 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; and

repeating 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,

wherein the bias vector is based on a direction defined between the first mean vector and the second mean vector;

in response to reaching the extremum of the bias score, linearly project each embedding vector of the plurality of embedding vectors along the bias vector to debias the plurality of embedding vectors to provide a plurality of debiased embedding vectors;

train a machine learning model of a fraud monitoring system based on the plurality of debiased embedding vectors;

receive an authorization request associated with a transaction; and

decline the transaction associated with the authorization request based on the machine learning model of the fraud monitoring system as trained based on the plurality of debiased embedding vectors.

7. The system of claim 6, wherein the plurality of embedding vectors are vector representations of merchant identity embeddings generated from customer transaction data.

8. The system of claim 6, wherein the bias score is calculated from a Word Embedding Association Test (WEAT).

9. The system of claim 6, wherein the bias score is calculated from an Embedding Coherence Test (ECT).

10. The system of claim 6, wherein the at least one 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.

11. 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;
determine a bias vector by:

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

modifying 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 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 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 the bias score associated with each second candidate vector;

replacing 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; and

repeating 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,

wherein the bias vector is based on a direction defined between the first mean vector and the second mean vector;

in response to reaching the extremum of the bias score, linearly project each embedding vector of the plurality of embedding vectors along the bias vector to debias the plurality of embedding vectors to provide a plurality of debiased embedding vectors;

train a machine learning model of a fraud monitoring system based on the plurality of debiased embedding vectors;

receive an authorization request associated with a transaction; and

decline the transaction associated with the authorization request based on the machine learning model of the fraud monitoring system as trained based on the plurality of debiased embedding vectors.

12. The computer program product of claim 11, wherein the plurality of embedding vectors are vector representations of merchant identity embeddings generated from customer transaction data.

13. The computer program product of claim 11, wherein the bias score is calculated from a Word Embedding Association Test (WEAT).

14. The computer program product of claim 11, wherein the bias score is calculated from an Embedding Coherence Test (ECT).

15. The computer program product of claim 11, 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.