Supplementary Materials

Table S1. Comparison of precision between SPPMI-based keyword expansion and five other baseline models on extracting never smoker-related keywords (Word co-occurrence, PMI vector, NPMI vector, PMI score, and NPMI score models). The values of d represent a number of singular values used in SVD.

# of keywords **Methods**	**Top 1**	**Top 5**	**Top 10**	**Top 20**
Word co-occurrence	**36.36%**	14.55%	8.18%	6.82%
PMI vector	27.27%	9.09%	5.45%	4.09%
NPMI vector	27.27%	10.91%	9.09%	5.00%
PMI score	18.18%	**16.36%**	**13.64%**	**11.36%**
NPMI score	18.18%	12.73%	12.73%	**11.36%**
SPPMI ($k = 1$)	27.27%	9.09%	5.45%	4.09%
SPPMI ($k = 5$)	**36.36%**	**16.36%**	9.09%	5.91%
SPPMI ($k = 15$)	**36.36%**	14.55%	8.18%	6.82%
SPPMI-SVD ($k = 1, d = 100$)	18.18%	3.64%	5.45%	3.64%
SPPMI-SVD ($k = 1, d = 500$)	27.27%	12.73%	7.27%	4.09%
SPPMI SVD ($k = 1, d = 1000$)	**36.36%**	9.09%	6.36%	3.18%
SPPMI-SVD ($k = 5, d = 100$)	18.18%	10.91%	5.45%	3.64%
SPPMI-SVD ($k = 5, d = 500$)	**36.36%**	12.73%	6.36%	3.64%
SPPMI- SVD ($k = 5, d = 1000$)	**36.36%**	12.73%	6.36%	3.18%
SPPMI SVD ($k = 15, d = 100$)	18.18%	12.73%	6.36%	4.55%
SPPMI- SVD ($k = 15, d = 500$)	27.27%	12.73%	7.27%	3.64%
SPPMI-SVD ($k = 15, d = 1000$)	**36.36%**	12.73%	7.27%	3.64%

Table S2. Comparison of precision between SPPMI-based keyword expansion and five other baseline models on extracting past smoker-related keywords (Word co-occurrence, PMI vector, NPMI vector, PMI score, and NPMI score models). The values of d represent a number of singular values used in SVD.

# of keywords **Methods**	**Top 1**	**Top 5**	**Top 10**	**Top 20**
Word co-occurrence	31.25%	42.50%	35.63%	**35.94%**
PMI vector	37.50%	41.25%	33.13%	31.25%
NPMI vector	37.50%	38.75%	32.50%	30.00%
PMI score	37.50%	38.75%	36.88%	35.63%
NPMI score	25.00%	42.50%	37.50%	33.75%
SPPMI $(k = 1)$	37.50%	41.25%	33.13%	31.25%
SPPMI $(k = 5)$	25.00%	32.50%	33.13%	30.63%
SPPMI $(k = 15)$	37.50%	33.75%	28.75%	28.44%
SPPMI-SVD $(k = 1, d = 100)$	56.25%	43.75%	**41.25%**	35.00%
SPPMI-SVD $(k = 1, d = 500)$	37.50%	**47.50%**	38.13%	35.31%
SPPMI SVD $(k = 1, d = 1000)$	43.75%	45.00%	33.13%	30.94%
SPPMI-SVD $(k = 5, d = 100)$	**68.75%**	35.00%	29.38%	27.19%
SPPMI-SVD $(k = 5, d = 500)$	25.00%	33.75%	31.25%	31.56%
SPPMI- SVD $(k = 5, d = 1000)$	25.00%	36.25%	35.63%	32.81%
SPPMI SVD $(k = 15, d = 100)$	62.50%	40.00%	30.00%	27.81%
SPPMI- SVD $(k = 15, d = 500)$	56.25%	36.25%	28.75%	28.13%
SPPMI-SVD $(k = 15, d = 1000)$	43.75%	32.50%	29.38%	29.06%

Table S3. Comparison of precision between SPPMI-based keyword expansion and five other baseline models on extracting current smoker-related keywords (Word co-occurrence, PMI vector, NPMI vector, PMI score, and NPMI score models). The values of d represent a number of singular values used in SVD.

# of keywords **Methods**	**Top 1**	**Top 5**	**Top 10**	**Top 20**
Word co-occurrence	43.48%	40.87%	37.39%	35.87%
PMI vector	52.17%	46.09%	42.17%	35.87%
NPMI vector	47.83%	45.22%	42.61%	36.96%
PMI score	47.83%	45.22%	48.26%	44.78%
NPMI score	52.17%	47.83%	49.13%	**45.43%**
SPPMI (k = 1)	52.17%	46.09%	42.17%	35.87%
SPPMI (k = 5)	47.83%	45.22%	41.74%	35.43%
SPPMI (k = 15)	34.78%	35.65%	34.78%	36.09%
SPPMI-SVD (k = 1, d = 100)	52.17%	**56.52%**	**52.17%**	42.39%
SPPMI-SVD (k = 1, d = 500)	60.87%	50.43%	46.96%	38.26%
SPPMI SVD (k = 1, d = 1000)	60.87%	48.70%	41.74%	41.09%
SPPMI-SVD (k = 5, d = 100)	65.22%	52.17%	46.52%	43.04%
SPPMI-SVD (k = 5, d = 500)	56.52%	48.70%	46.52%	41.52%
SPPMI- SVD (k = 5, d = 1000)	56.52%	43.48%	43.48%	41.09%
SPPMI SVD (k = 15, d = 100)	**69.57%**	53.91%	48.70%	45.00%
SPPMI- SVD (k = 15, d = 500)	43.48%	39.13%	41.30%	42.39%
SPPMI-SVD (k = 15, d = 1000)	47.83%	43.48%	43.48%	40.00%

Table S4. Comparison of never smoker classification accuracy. All accuracies are reported in F1 scores.

# of extracted keyword / **Methods**	**1**	**5**	**10**	**20**
Bag of Words	91.57%			
SPPMI (k = 1)	92.37%	91.63%	91.86%	92.37%
SPPMI (k = 5)	91.33%	90.61%	91.29%	92.21%
SPPMI (k = 15)	91.14%	90.61%	91.51%	91.89%
SPPMI-SVD (k = 1, d = 100)	92.47%	91.10%	92.08%	90.64%
SPPMI-SVD (k = 1, d = 500)	91.37%	90.87%	91.29%	92.60%
SPPMI SVD (k = 1, d = 1000)	91.14%	91.40%	92.63%	**93.25%**
SPPMI-SVD (k = 5, d = 100)	91.33%	91.48%	91.82%	91.63%
SPPMI-SVD (k = 5, d = 500)	91.14%	91.63%	92.05%	91.82%
SPPMI- SVD (k = 5, d = 1000)	91.56%	91.56%	92.50%	91.21%
SPPMI SVD (k = 15, d = 100)	90.95%	91.79%	91.29%	92.80%
SPPMI- SVD (k = 15, d = 500)	90.95%	91.60%	91.82%	90.83%
SPPMI-SVD (k = 15, d = 1000)	91.60%	90.99%	91.40%	91.79%

Table S5. Comparison of past smoker classification accuracy. All accuracies are reported in F1 scores.

# of extracted keyword / **Methods**	**1**	**5**	**10**	**20**
Bag of Words	89.74%			
SPPMI ($k = 1$)	87.92%	87.66%	88.72%	88.66%
SPPMI ($k = 5$)	88.72%	86.01%	89.46%	90.13%
SPPMI ($k = 15$)	88.66%	88.43%	87.14%	88.48%
SPPMI-SVD ($k = 1, d = 100$)	88.14%	86.67%	86.53%	89.12%
SPPMI-SVD ($k = 1, d = 500$)	89.06%	87.96%	88.14%	90.21%
SPPMI SVD ($k = 1, d = 1000$)	88.02%	89.46%	90.54%	**90.72%**
SPPMI-SVD ($k = 5, d = 100$)	88.60%	89.06%	86.38%	89.29%
SPPMI-SVD ($k = 5, d = 500$)	88.48%	89.00%	88.49%	90.18%
SPPMI- SVD ($k = 5, d = 1000$)	88.37%	88.61%	90.08%	90.31%
SPPMI SVD ($k = 15, d = 100$)	88.89%	87.34%	86.45%	88.44%
SPPMI- SVD ($k = 15, d = 500$)	87.66%	87.40%	88.83%	88.89%
SPPMI-SVD ($k = 15, d = 1000$)	88.54%	88.49%	88.49%	89.17%

Table S6. Comparison of current smoker classification accuracy. All accuracies are reported in F1 scores.

# of extracted keyword **Methods**	**1**	**5**	**10**	**20**
Bag of Words	77.88%			
SPPMI ($k = 1$)	80.18%	81.19%	82.52%	80.95%
SPPMI ($k = 5$)	84.91%	78.64%	84.21%	84.62%
SPPMI ($k = 15$)	82.19%	83.96%	83.18%	84.51%
SPPMI-SVD ($k = 1, d = 100$)	80.37%	80.58%	80.75%	83.02%
SPPMI-SVD ($k = 1, d = 500$)	82.88%	84.11%	82.86%	84.36%
SPPMI SVD ($k = 1, d = 1000$)	82.19%	83.57%	84.91%	**86.92%**
SPPMI-SVD ($k = 5, d = 100$)	81.45%	84.51%	79.43%	82.52%
SPPMI-SVD ($k = 5, d = 500$)	82.51%	82.46%	83.41%	84.47%
SPPMI- SVD ($k = 5, d = 1000$)	81.65%	82.46%	85.02%	83.25%
SPPMI SVD ($k = 15, d = 100$)	83.57%	81.13%	80.58%	83.33%
SPPMI- SVD ($k = 15, d = 500$)	81.25%	81.90%	82.30%	81.55%
SPPMI-SVD ($k = 15, d = 1000$)	82.73%	82.76%	83.81%	81.73%

Table S7. Comparison of unknown smoking status classification accuracy. All accuracies are reported in F1 scores.

# of extracted keyword Methods	1	5	10	20
Bag of Words	93.05%			
SPPMI (k = 1)	93.81%	94.68%	94.45%	94.61%
SPPMI (k = 5)	93.96%	94.31%	94.54%	94.31%
SPPMI (k = 15)	93.66%	94.79%	94.55%	94.32%
SPPMI-SVD (k = 1, d = 100)	94.13%	94.18%	94.18%	94.42%
SPPMI-SVD (k = 1, d = 500)	93.66%	94.06%	94.54%	94.35%
SPPMI SVD (k = 1, d = 1000)	93.45%	93.68%	94.55%	**95.06%**
SPPMI-SVD (k = 5, d = 100)	93.55%	94.31%	94.45%	94.57%
SPPMI-SVD (k = 5, d = 500)	93.43%	94.54%	94.54%	94.54%
SPPMI- SVD (k = 5, d = 1000)	93.43%	93.80%	94.29%	94.42%
SPPMI SVD (k = 15, d = 100)	94.20%	93.84%	94.18%	94.83%
SPPMI- SVD (k = 15, d = 500)	93.55%	93.71%	94.54%	94.78%
SPPMI-SVD (k = 15, d = 1000)	93.55%	93.74%	94.55%	94.54%