基于中国社会评价词表的语义相似度计算与词表扩充

朱贺，李璐璐，陈庆荣，胡传鹏

南京师范大学心理学院

**摘要**

由于社会评价层面的词汇在日常生活中的使用较为丰富，现有的中文社会评价词表在词汇数量上仍存在进一步扩大的空间，以提升词表的多样性。本文基于中文社会评价词表(The Chinese Social Evaluatige Word List, Li, Chen, & Chuan-Peng, 2023)，对五个社会评价维度中的1,133个中文词汇进行了扩充。本文采用Word2Vec词嵌入方法对词表进行计算扩充，并将其与GloVe词向量库所得出的结果进行交互验证，对扩展词表的信度与效度进行检验。研究结果表明，扩展后的词表在一定程度上保留了原始词表的语义信息，并在外部验证中表现出较高的可靠性和有效性，可以为社会评价文本分析提供一定的数据支持。

*关键词：*词表扩充，词嵌入，社会评价词表

基于中国社会评价词表的语义相似度计算与词表扩充

**Introduction**

社会评价词表的全面性与可解释性对社会评价等研究具有重要意义，但目前的词表规模较小，收入的词汇数量仍然不够充足。社会评价广泛地发生在生活中，构建综合性的标准社会评价词表是研究、量化社会评价的重要手段之一。现有中文社会评价词表在词汇数量方面仍有欠缺，这局限了词表对社会评价的全面反映，在应用上难以满足多样化的社会评价场景。

词嵌入这一方法广泛应用于计算扩展词表的相关研究当中。其或作为辅助，或成为主导的扩展算法。Nicolas等人( 2021)使用Word2Vec作为文本处理工具，在扩展词表的主要工作中采用WordNet词网进行语义相似度的计算，从而达到扩充刻板印象词表的目的，并非采取静态词嵌入的方法进行词表扩展。相比之下，S.Wang( 2023)使用Word2Vec模型和MacBERT模型提取中文词语的词表征并进行语义评分，直接采用新华社新闻报道语料库作为建立词向量模型所需的语料库。类似地，李晓(2017)基于中文维基百科语料作构建词向量模型，陆峰(2017)则使用爬取的电商评价语料库以完成词嵌入工作。上述三项研究中均采用Word2Vec作为词嵌入方法，基于语料库嵌入词向量，得到词向量模型，并通过计算词语间的余弦值来衡量相关性，依据一定标准筛选出余弦值较大的词汇组成扩展词表。在备选词的筛选标准方面，杨阳(2014)具体提出了选用备选词的筛选标准为采用语义相似度大于0.8的词汇，但并未明确表明选择余弦相似度（cosine similarity）为0.8作为筛选阈值的依据。

本文将采用词嵌入这一方法，扩展已有的中文社会评价词表。基于The Chinese Social Evaluative Word List（中文社会评价词表），采取两种静态词嵌入方法对词表进行向量化，进而通过对词向量之间的语义相似度进行计算并筛选备选词，实现对现有中文社会评价词表的进一步扩充。

**Methods**

本文采取词嵌入（word embedding）的方法，使用Word2Vec (Word to Vector, Mikolov et al., 2013)和GloVe (Global Vectors for Word Representation, Pennington et al., 2014)两种上下文不敏感性模型，进而对计算得出的扩展词表进行交互验证。

Word2Vec、GloVe两种方法同属于静态词嵌入类别。静态词嵌入是指，将一个词在整个语料库中的所有上下文信息都聚合、压缩到一个向量表示中，得到的是固定的、不随词汇所在特定语境中的上下文变化的词向量(Bao et al., 2023)。

Word2Vec是Google于2013年开源的词向量工具（[https://code.google.com/archive/p/word2vec/](file:///C:\\Users\\jiapi\\Desktop\\%5bWord2Vec%20Project%20Archive%5d(https:\\code.google.com\\archive\\p\\word2vec\\))），它利用深度学习的思想，可以通过训练，将语料中的词语映射为K维向量空间中的向量。“该模型通过学习训练语料获得词向量和概率密度函数，词向量是多维实数向量，向量中包含了自然语言中的语义和语法关系，词向量之间余弦距离的大小代表了词语之间关系的远近。”（熊富林，2015）余弦距离的大小即为两个词向量在向量空间中夹角的余弦值，可以衡量词语之间的语义相似度。

GloVe是斯坦福大学于2014年发布的词向量工具（<https://nlp.stanford.edu/projects/glove/>），它根据语料库中全局的词与词共现统计数据进行训练，进而获取词的向量表征。GloVe通过计算两个词向量之间的欧式距离或余弦相似度来衡量词在语义上的相似性，可以根据这一指标得出相应词向量的最近临近词。此外，GloVe还可以计算两个词并列时它们的向量差异，如“男人-女人”、“国王-王后”等词向量组(Pennington et al., 2014)。

这两种静态词嵌入方法都常用于计算词与词之间的语义相似度，从而达到聚类文本、扩展词典、情感分析等作用。它们都采用无监督方法计算，并在语义相似度度量上有一致性，均可通过余弦相似度或欧式距离找到相关词语，从而实现词表扩展的目标。因此，使用Word2Vec词向量库作为惟一的评估信息基础存在一定局限性。基于GloVe词向量库进行分析，并将两者在词向量空间中的嵌入结果进行对比，可以进一步提高词表的稳健性，这有助于更精确地验证扩展词表的有效性和准确性(Curto et al., 2024)。

***扩展词表实验设计***

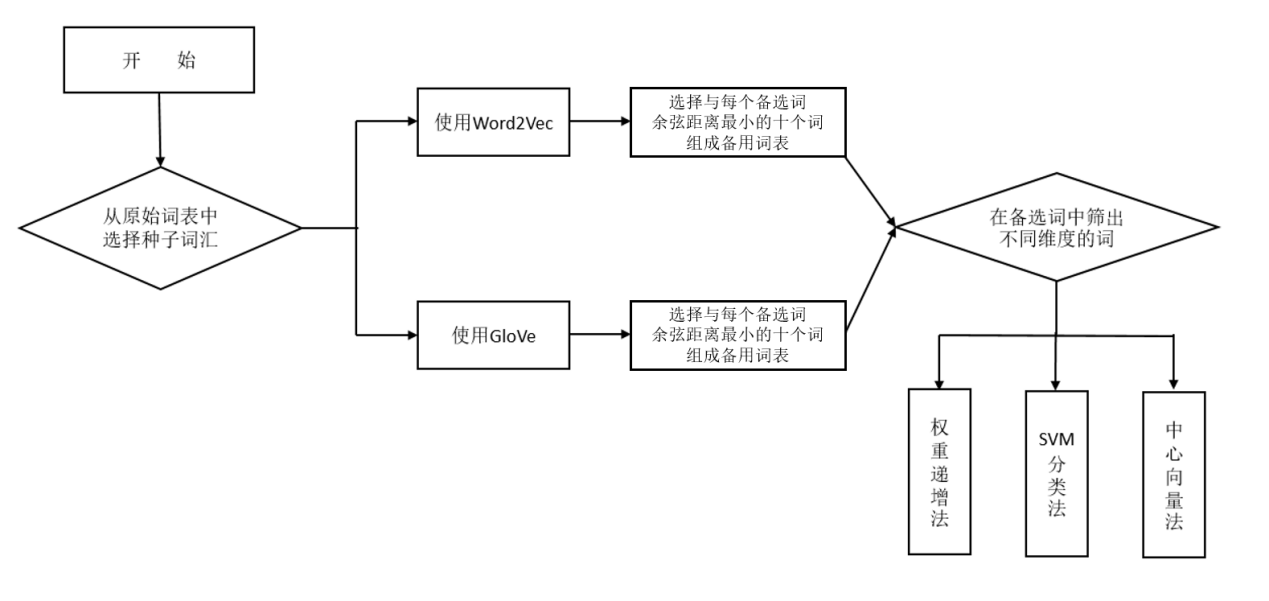


Figure 1: 词表扩展流程图

本研究通过词嵌入方法扩展现有的社会评价词表。用于计算词汇语义相似性、相关性的主要技术是使用来自于词嵌入的Word2Vec模型(Jatnika et al., 2019)。计算词与词之间的语义联系，一种方法是直接计算词向量的余弦相似度或距离。具体而言，两个词向量在空间中夹角的余弦值即余弦相似度，这一数值可以衡量两个词语之间的语义关联性（semantic relatedness）(Bao et al., 2023)。余弦相似度的计算公式如下：

用1减去余弦相似度，则得出余弦距离（cosine distance）。因此，若词语之间在语义上较为相似时，那它们之间的余弦相似度会更大、余弦距离更小，并且在向量空间的位置也更接近。

基于此原理，本文将现有的中文社会评价词表总体作为一个种子词汇集，使用腾讯发布的AILab词向量(Song et al., 2018)作为词向量库。参照中文道德词典的构建方法(H. Wang et al., 2020)，在使用Word2Vec构建扩展词表的过程中，选用向量库中与每个种子词汇余弦距离最小的十个词作为备选词汇。之后，在生成的新词汇群中去除重叠部分，形成扩展词表的备选词群。

Table 1: 词向量扩展词汇示例

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 种子词 | 扩展 | | | | |
| 首富 | 富足的生活 | 身价飙升 | 鲜花着锦 | 很气派 | 土豪们 |
| 富于 | 家境殷实 | 下流 | 飞黄腾达 | 家资 |
| 交际花 | 随性不羁 | 不会主动 | 咋咋呼呼 | 性格直爽 | 嘻嘻哈哈 |
| 狂放 | 关系亲近 | 十分豪爽 | 口直心快 | 口才了得 |
| 无能 | 百无一用 | 一无可取 | 花架子 | 害怕困难 | 低智 |
| 诡计多端 | 智商有问题 | 邪门儿 | 假把式 | 没用的家伙 |

使用余弦相似度或余弦距离得到的词汇虽与种子词汇的相似度较高，但并不代表其语义一定相同或相似。词相似度衡量的是语义联系的绝对值，既不必然表示联系的正、负方向，也难以直接区分同义词和反义词(Bao et al., 2023)。因此，通过Word2Vec模型计算余弦最短距离而生成的新词，其情感极性，即消极性质（negative）或积极性质（positive）不一定与种子词相同。以积极性质的词汇为种子词所扩展得到的词，也有可能具有消极性质。因此，结束备选词的生成、去重工作之后，还需进一步判断词的情感极性。

本文选择中心向量法作为判断词汇情感极性的方法。中心向量法是指人工选取语料中正向、负向情感极性评价词汇，将标注为同一极性的评价词对应词语向量各维度相加后取平均值，从而计算出这一极性类别评价词的中心向量，之后通过此中心向量与语料中未标注词语的词向量进行余弦相似度的计算(杨阳 et al., 2014)。本文将种子词表中积极性质的词汇、消极性质的词汇各作为一类，分别计算词向量的平均值，将其作为各自的中心向量。之后获取备选词汇的向量，计算各备选词与两类中心向量的余弦相似度。若备选词与积极词汇中心向量的余弦相似度大于消极词汇中心向量，则该词被判定为积极性质；反之，则被判定为消极性质。

生成备选词并对其进行去重、确定情感极性并进行人工筛选后，将备选词作为中文社会评价词表的扩展词表。

**信效度检验**

***相关性检验：皮尔逊相关系数（Pearson Correlation coefficient）***

皮尔逊相关系数用于衡量两个变量之间线性关系的强度和方向，可以取-1到+1之间的值，值越大表示相关性越强。+1表示变量与递增关系完全线性相关的，-1则表示变量与递减关系完全线性相关(Gogtay & Thatte, n.d.)。如果相关系数大于0.8，则认为词之间是强相关的关系；若相关系数小于0.5，则一般认为是弱相关(Bolboacă & Jäntschi, 2006).。在词嵌入相关的研究中，常用于对Word2Vec所生成词向量的相关性判断中(Jatnika et al., 2019)。

本文使用皮尔逊相关系数进行扩展词表与社会评价词表的相关性检验。分别计算社会评价词表、扩展词表中各维度积极、消极词汇的词向量平均值，得出中心向量后，比较两个中心向量之间的皮尔逊相关系数。若相关系数超过0.8，则表明该维度的正负情感极性词汇之间具有显著相关性；若所有维度均达到此标准，则证明扩展词表与社会评价词表具有一致性。

***有效性检验：基于社会评价词表的分类任务***

由于缺乏与该社会评价词表维度划分相似的中文词表，本文难以通过与其它现有词表对比分析以检验扩展词表的有效性，因此采用机器学习方法，基于社会评价词表进行分类任务，以检验这一扩展词表是否有效可行。

本文通过机器学习模型进行的分类任务主要针对于识别词汇所在维度。从腾讯AILab词向量库中提取种子词表和扩展词表的向量表达作为训练数据，使用扩展词汇的五个维度信息，即外貌、社会经济地位、社交能力、能力、道德作为划分标签，构建分类任务所使用的数据集。在筛去不具有词向量的词汇后，最终数据集的大小为9,693个词。。数据集完成后，进行数据清洗与整理，将词向量部分的数据格式整理为以逗号为分隔的列表形式，将维度标签中的外貌、社会经济地位、社交能力、能力、道德分别赋予0、1、2、3、4的数值标签。将数据集以8:2的比例划分为训练集与测试集。

确立训练集、验证集确立后，构建并训练分类器。在分类任务的模型选择方面，本文选择采用三种常见的分类算法：支持向量机（Support Vector Machine，SVM）、逻辑回归（Logistic Regression，LR）和极端梯度提升（Extreme Gradient Boosting，XGBoost），以进行对比实验。

将训练集输入模型，进行训练。在分类器训练完毕后，使用测试集，对维度标签进行预测。针对所得预测结果，使用F1值（F1-Score）和混淆矩阵（Confusion Matrix）作为模型表现和性能对比的评估工具。其中，F1值综合衡量模型的精确率与召回率，更平衡地体现模型分类表现；混沌矩阵则提供分类结果的详细分布，以展示模型在不同维度的预测情况。F1值最高的模型，则表示对词汇维度的分类预测效果最佳。若F1值大于0.8，即表示模型能够达到较高的准确率，表明扩展词表与社会评价词表具有一致性，从而验证了扩展词表的有效性。

**结果**

***扩展词表实验结果***

本文通过词嵌入的方法，基于余弦距离的大小，共得到8,590个扩展词。

在外貌这一维度中，原始词表共有256个词汇。根据Word2Vec给出的词向量，计算余弦最短距离进行扩展并去除重叠词后，得到备选词汇1,982个。通过中心向量法进行情感极性分类，得到积极性质的新词1,118个，消极性质的新词864个。

在社会经济地位这一维度中，原词表共有113个词汇。其中积极性质的词汇41个，消极性质的词汇72个。在扩展及去除重叠词后，得到备选词汇911个。对新词进行情感极性分类，得到382个积极性质词汇，529个消极性质词汇。

在社交能力这一维度中，原始词表含有111个词汇。其中积极性质的词汇64个，消极性质的词汇47个。经词向量扩展及去重后，得到备选词汇855个。对备选词汇进行情感极性分类，得到积极性质的新词458个，消极性质的新词397个。

在能力这一维度中，原始词表中有284个词汇。其中积极性质的词汇206个，消极性质的词汇78个。扩展及去重后，得到备选词汇2,281个。对备选词汇进行情感极性分类，得到积极性质的新词1,583个，消极性质的新词698个。

在道德这一维度中，原始词表含有369个词汇。其中积极性质的词汇176个，消极性质的词汇193个。扩展及去重后，得到备选词汇2,559个。在进行基于中心向量法的情感极性分类后，得到1,224个积极性质词汇，1,335个消极性质词汇。

扩展词表的有效新词数量为8,588个。其中，五个维度的积极性质词汇有4,765个，消极性质词汇有3,823个。相较而言，积极词汇的规模比起消极词汇的更大。

***相关性检验实验结果***

在信效度检验方面，本文在进行相关性检验时，使用皮尔逊相关系数作为衡量指标，以计算扩展词表与种子词表的相似度。结果表明，外貌维度中，积极性质的种子词与扩展词表之间的皮尔逊相关系数达到0.9694，消极性质的种子词与扩展词表之间的皮尔逊相关系数达到0.9778。社会经济地位维度中，积极性质的种子词与扩展词表之间的皮尔逊相关系数达到0.9690，消极性质的种子词与扩展词表之间的皮尔逊相关系数达到0.9749。社交能力维度中，积极性质的种子词与扩展词表之间的皮尔逊相关系数达到0.9756，消极性质的种子词与扩展词表之间的皮尔逊相关系数达到0.9830。能力维度中，积极性质的种子词与扩展词表之间的皮尔逊相关系数达到0.9765，消极性质的种子词与扩展词表之间的皮尔逊相关系数达到0.9831。道德维度中，积极性质的种子词与扩展词表之间的皮尔逊相关系数达到0.9776，消极性质的种子词与扩展词表之间的皮尔逊相关系数达到0.9869。这证明具备五个维度的扩展词表与社会评价词表之间均有显著的相关性。

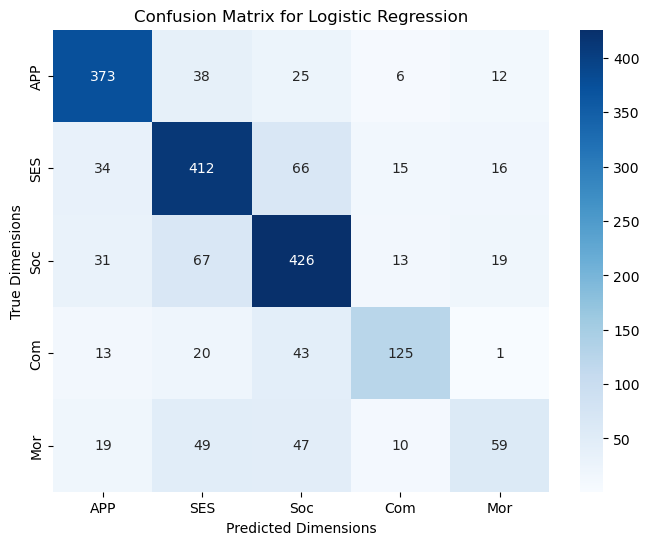
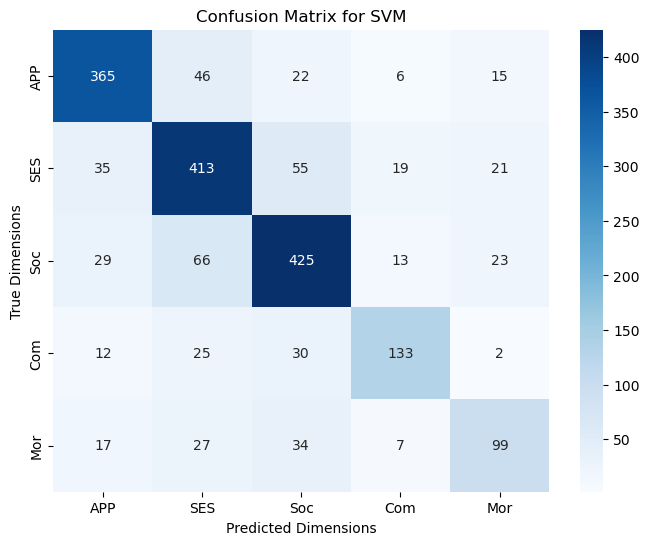
***有效性检验实验结果***

在有效性方面，本文采用机器学习方法，对含有词汇、对应词向量进而维度标签的数据集进行分类任务，预测数据集中验证集部分的词汇维度。所使用的SVM、LR、XGBoost三种模型，其F1值和混淆矩阵结果如下：

Table 2: SVM、RF、XGBoost的实验结果对比

|  |  |  |  |
| --- | --- | --- | --- |
| **模型** | **Precision** | **Recall** | **F1-Score** |
| SVM | 0.739 | 0.740 | 0.738 |
| LR | 0.713 | 0.719 | 0.712 |
| XGBoost | 0.792 | 0.792 | 0.791 |

分类结果显示，XGBoost模型的F1值达到0.791，证明该模型可以较好学习社会评价词汇的词向量特征，以此为数据基础预测未知词汇的维度标签。



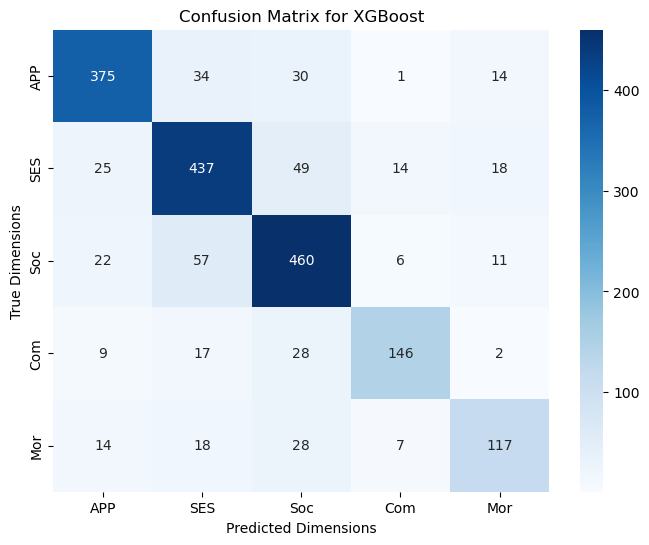


Figure 2: SVM、RF、XGBoost的混淆矩阵热力图

由三种机器学习模型所生成的混淆矩阵热力图，可以看出，社交能力维度的分类预测表现最为准确，模型对该类别的分类正确率较高、误分类数量较少。其次是对社会经济、外貌维度的预测。

道德维度的预测表现则不理想。在三个模型中，道德维度的误分类情况表现得较为相似，说明该维度的扩展词表与原词表，其词向量特征未能被显著区分。这表示，道德维度的扩展词表相较于其他维度，其在特征上具有重叠性或模糊性，使得模型在分类预测时表现较弱。

综上所述，该结果验证了扩展词表的有效性，表明扩展词表在社会评价的五个维度方面与社会评价词表具有较高的一致性。其中，社交能力、社会经济、外貌维度的扩展词表具有更高的有效性、精确性，分类模型对其的学习效果优于道德维度。

**讨论**

在建立词向量模型方面，使用腾讯AILab的公开词嵌入语料库，在这一语料库所涵盖的范围内找出语义相似度最接近的词汇。H. Wang(H. Wang et al., 2020)即使用了这一词向量库。依据这一研究，确立将与原始词表中每个词余弦距离最小的十个词作为扩展词表备选词的扩展标准。但十个词是否能够作为最合适的扩展标准，还没有一个具体的确立指标。

扩展词表中不同维度之间的词汇有可能出现维度重叠现象。在经过算法分类后，还需要人工进行进一步的查筛。

尽管皮尔逊相关系数的结果呈现相关性，但由于扩展词表和社会评价词表的词向量都通过腾讯AILab词嵌入语料库与Word2Vec生成，两者之间的词向量一定程度上已具备来自同一模型所导致的相似性。相关性高的结果可能在一定程度上反映出该词嵌入算法本身的性质，而不是扩展词表与社会评价词表之间的语义关联。

**结论**

本文进一步扩展了中文评价词表，通过词嵌入模型增加了8,588个中文词汇。使用Word2Vec这一词嵌入方法，将社会评价词表的规模由1,133个扩展至9,721个。

本文利用词嵌入对中文评价词表进行了计算扩展，并验证了扩展词表的信效度。研究结果显示，扩展后的词表在有效性和可靠性方面都表现良好，可供未来词汇表的扩展研究参考。

**参考文献**

Bao H.-W.-S., Wang Z.-X., Cheng X., Su Z., Yang Y., Zhang G.-Y., Wang B., & Cai H.-J. (2023). Using word embeddings to investigate human psychology: Methods and applications. *Advances in Psychological Science*, *31*(6), 887. https://doi.org/10.3724/SP.J.1042.2023.00887

Bolboacă, S.-D., & Jäntschi, L. (2006). *Pearson versus Spearman, Kendall’s Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds*. *9*.

Curto, G., Jojoa Acosta, M. F., Comim, F., & Garcia-Zapirain, B. (2024). Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings. *AI & SOCIETY*, *39*(2), 617–632. https://doi.org/10.1007/s00146-022-01494-z

Gogtay, N., & Thatte, U. (n.d.). *Principles of Correlation Analysis*.

Jatnika, D., Bijaksana, M. A., & Suryani, A. A. (2019). Word2Vec Model Analysis for Semantic Similarities in English Words. *Procedia Computer Science*, *157*, 160–167. https://doi.org/10.1016/j.procs.2019.08.153

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. http://arxiv.org/abs/1301.3781

Nicolas, G., Bai, X., & Fiske, S. T. (2021). Comprehensive stereotype content dictionaries using a semi‐automated method. *European Journal of Social Psychology*, *51*(1), 178–196. https://doi.org/10.1002/ejsp.2724

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Song, Y., Shi, S., Li, J., & Zhang, H. (2018). Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 175–180). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2028

Wang, H., Liu, C., & Yu, D. (2020). 面向人工智能伦理计算的中文道德词典构建方法研究(Construction of a Chinese Moral Dictionary for Artificial Intelligence Ethical Computing). In M. Sun, S. Li, Y. Zhang, & Y. Liu (Eds.), *Proceedings of the 19th Chinese National Conference on Computational Linguistics* (pp. 539–549). Chinese Information Processing Society of China. https://aclanthology.org/2020.ccl-1.50

Wang, S., Zhang, Y., Shi, W., Zhang, G., Zhang, J., Lin, N., & Zong, C. (2023). A large dataset of semantic ratings and its computational extension. *Scientific Data*, *10*(1), Article 1. https://doi.org/10.1038/s41597-023-01995-6

李晓, 解辉, & 李立杰. (2017). 基于Word2vec的句子语义相似度计算研究. 计算机科学, *44*(9), 256–260.

杨阳, 刘龙飞, 魏现辉, & 林鸿飞. (2014). 基于词向量的情感新词发现方法. 山东大学学报(理学版), *49*(11), 51–58. https://doi.org/10.6040/j.issn.1671-9352.3.2014.255

陆峰. (2017). 基于word2vec扩充情感词典的商品评论倾向分析. 电脑知识与技术, *13*(5), 143-145+159. https://doi.org/10.14004/j.cnki.ckt.2017.0623