Performance Evaluation of Word Embedding Algorithms

1. Introduction

This study compares three word-embedding approaches—Skip-gram (Word2Vec), GloVe, and SPPMI–SVD—trained on a normalized edition of Pride and Prejudice under a uniform preprocessing and training protocol. Models are tuned via compact hyperparameter grids and evaluated with a shared intrinsic criterion: for a fixed set of 8 thematic probes, we compute cosine neighbourhoods and summarize quality by the mean similarity of each probe's top 5 neighbours (self-excluded). Qualitative inspection (e.g., gender, class, marriage discourse) complements the metric, clarifying strengths and limitations of predictive, count-based, and hybrid methods on a literary corpus.

2. Explanation of Algorithm

We compare three widely used word-embedding algorithms under a unified preprocessing and evaluation pipeline: a neural-network–based model (Skip-gram/Word2Vec), a word-matrix–based model (SPPMI–SVD), and a hybrid model (GloVe). Below we summarize the training objective, learning signal, and typical properties of each approach.

2.1 Neural-network–based methods

2.1.1 Skip-gram (Word2Vec)

Skip-gram is a predictive model that learns word vectors by maximizing the probability of context words given a target word within a fixed window. Each training instance consists of a centre token and its surrounding tokens; the model updates the embedding of the centre word so that it becomes a good predictor of its neighbours. Efficient training relies on techniques such as negative sampling (or hierarchical SoftMax), which replace the full SoftMax with a lightweight objective. Because updates are driven by many local contexts, Skip-gram is effective at capturing syntactic frames and semantic regularities and tends to perform well for infrequent words that still appear in informative contexts. Typical hyperparameters include vector dimensionality, window size, minimum frequency cutoff, and number of epochs.

2.2 Word-matrix–based methods

2.2.1 SPPMI–SVD

SPPMI–SVD is a count-based approach that starts from a word–context co-occurrence matrix built over a sliding window. Raw counts are converted to Pointwise Mutual Information (PMI) to emphasize informative associations rather than sheer frequency. To avoid negative values and to down-weight ubiquitous events, one applies a shifted positive PMI transformation,

$$\mathrm{SPPMI}(w, c) = \max(\mathrm{PMI}(w, c) - \log k, 0),$$

where $k > 1$ controls the shift. The resulting matrix is then factorized with Singular Value Decomposition (SVD), and the low-rank factors (for example U√S) serve as dense word embeddings. This method is conceptually simple and fast once the matrix is built, and it often highlights stable collocations and topical associations. Its main practical cost is memory for constructing and storing the co-occurrence matrix; performance is sensitive to vocabulary cutoffs, window size, and the choice of k and rank.

2.3 Hybrid methods

2.3.1 GloVe (Global Vectors)

GloVe combines the strengths of global co-occurrence statistics and local context learning. Like SPPMI–SVD, it begins from word–context co-occurrence counts, but instead of factorizing PMI it optimizes a weighted least-squares regression so that the dot product of word and context vectors approximates the log co-occurrence. A weighting function (with parameters such as x_{max} and α) down-weights very frequent pairs and ignores extremely rare ones, improving robustness. The key intuition is that ratios of co-occurrence probabilities encode semantic relations (for example, differences in ratios for ice vs. steam with respect to solid and gas). In practice, GloVe produces high-quality embeddings that balance global corpus statistics with local contextual information, and trains efficiently with stochastic updates.

3. Data Pre-processing

We construct a clean, sentence-segmented version of Pride and Prejudice (Project Gutenberg 1342). Boilerplate is removed via the canonical START/END markers. Sentences are tokenized (Punkt), words are lowercased and stripped of non-alphabetic characters; single letters are dropped. We remove English stop words and a curated set of high-frequency function words, then apply WordNet lemmatization (noun, then verb). Empty tokens are removed; empty sentences are skipped. The pre-processed corpus is written to corpus_tokens.txt (one sentence per line, space-separated). A fixed random seed (42) ensures reproducibility.

4. Model Training and Visualization

4.1 Model Training

We tune hyperparameters per model (e.g., window, dimensionality, weighting/regularization) under a consistent protocol (fixed seed and stable training settings). Evaluation uses the same intrinsic objective for all models: for a fixed probe set—{rich, poor, marriage, love, darcy, elizabeth, man, woman}—we L2-normalize embeddings, retrieve each probe's top-5 cosine neighbours (excluding the probe), and average the cosine scores across probes. These probes were chosen to reflect the corpus's central social dimensions: class dynamics via rich/poor, gender norms via man/woman, courtship/affect via marriage/love, and character-anchored discourse via

Darcy/Elizabeth. Each term appears with sufficient frequency after preprocessing, ensuring stable estimates. The highest-scoring configuration is retrained on the full corpus to produce final embeddings and neighbour summaries.

4.1.1 Skip-gram (Word2Vec).

Grid over window {2, 5, 10} and dimension {50, 100, 200}; train models, evaluate with the top-5-neighbor mean-cosine metric, select the best, retrain, and export top-5 neighbours.

4.1.2 GloVe (text2vec).

Grid over rank {50, 100, 150}, x_{max} {5, 10, 15}, window {3, 5, 8}, and iterations {30, 50, 100}. For each setting, build a TCM, train GloVe, and evaluate with the same top-5-neighbor mean-cosine metric. Select the best, keep its embeddings, and export top 5 neighbours.

4.1.3 SPPMI-SVD.

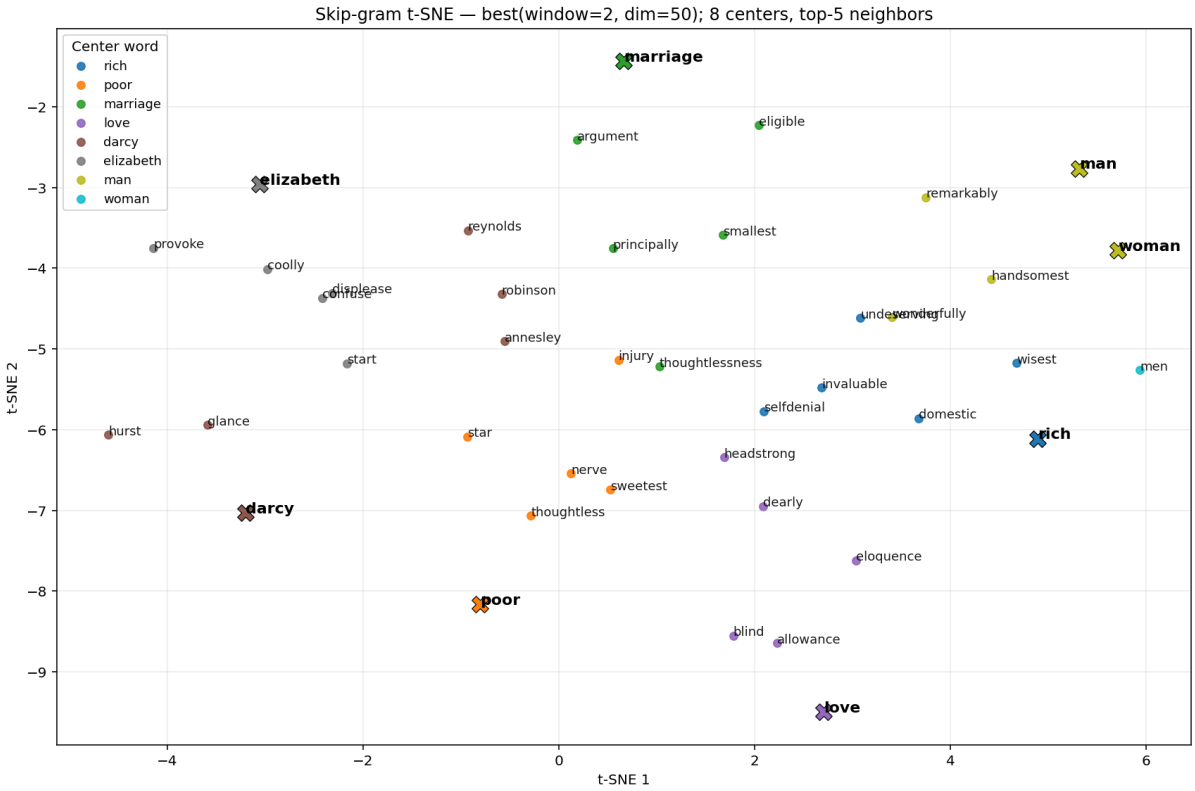
Grid over window, shift k, and rank. Build co-occurrence, compute SPPMI, factorize with SVD to obtain $U\sqrt{S}$, and evaluate with the same top-5-neighbor mean-cosine metric. Select the best, recompute with those settings, and export neighbours/embeddings.

4.2 Qualitative analysis of nearest neighbors within models

We use the same themed probes across methods (limited to those surviving frequency/cap filters) and inspect each probe's top 5 cosine neighbours for semantic/role alignment, syntactic compatibility (e.g., honorific + surname), and topical coherence (households, estates, courtship). Cosine magnitudes are interpreted within a model; cross-model numbers are not compared directly.

4.2.1 Skip-gram (Word2Vec).

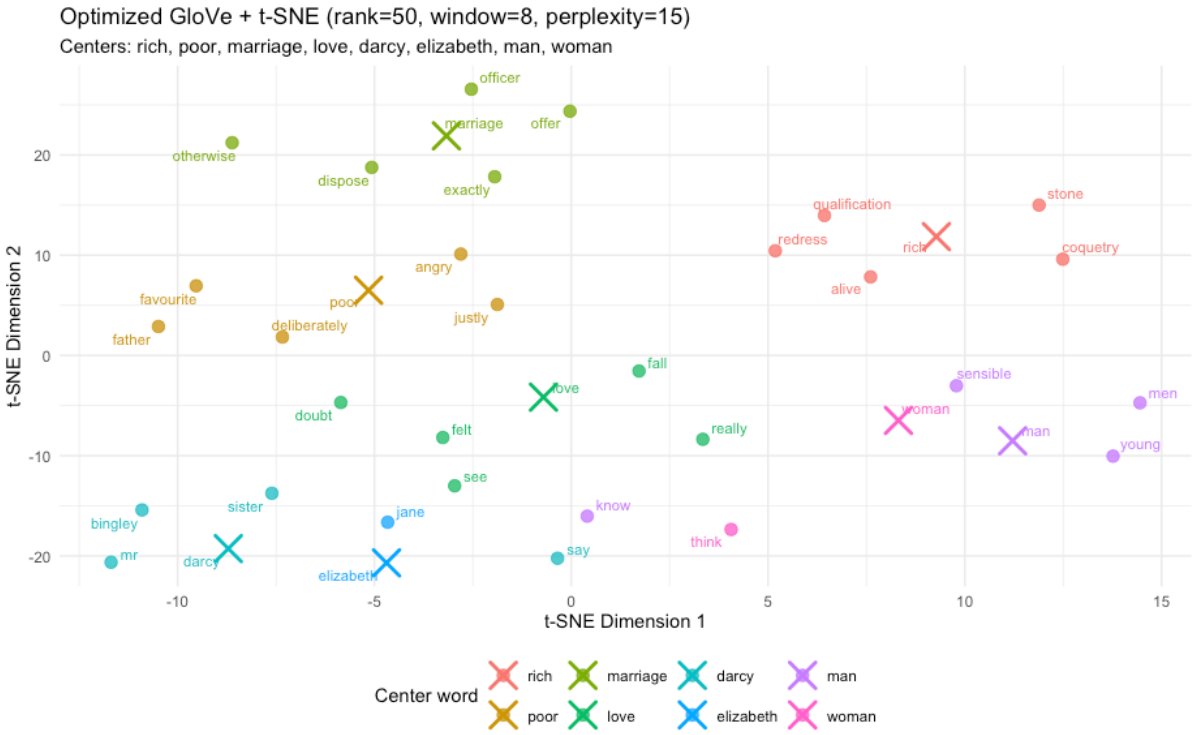
Overall, the neighborhoods are largely consistent with an Austen-era register: marriage aligns with procedural/evaluative lexis (argument, eligible, thoughtlessness), love captures idiom and affect (dearly, blind, headstrong), Darcy links to household/staff names (Annesley, Reynolds, Hurst, Robinson), Elizabeth clusters with stance/affect (coolly, displease, provoke), and man↔woman co-occur with appearance modifiers (handsomest, wonderfully, remarkably). Class signals appear with rich taking moral/domestic associates and poor leaning toward hardship/administration. Red flags include the broad modifier undeserving attaching to rich/man/woman, weakly thematic items for poor (star, sweetest, nerve, injury), and generic action words (glance, start, confuse/provoke) that suggest a corpus-wide stylistic axis rather than tight semantics.



4.2.2 GloVe (text2vec).

Overall, the neighborhoods are broadly plausible for an Austen register. Marriage is framed transactionally and procedurally—offer, dispose, exactly, officer, otherwise—consistent with courtship logistics. Love mixes affect and cognition—fall, felt, doubt, see, really. Characters are coherent: Darcy links to the social graph and dialogue (mr, elizabeth, sister, bingley, say), and Elizabeth to kin and speech (jane, sister, mr, darcy, say). Gender terms show asymmetry: man pairs with status/epistemic descriptors (sensible, know) plus young, while woman leans relational/attitudinal (man/men, think, really, young), reflecting patriarchal framing. Class signals are weaker: rich → qualification/stone/alive/redress/coquetry and poor → favourite/father/justly/deliberately/angry read as moral or stylistic rather than socioeconomic. Red flags include very high darcy → mr (0.90), generic function/stance words (say, really, think, know), and questionable class associates, suggesting a corpus-wide

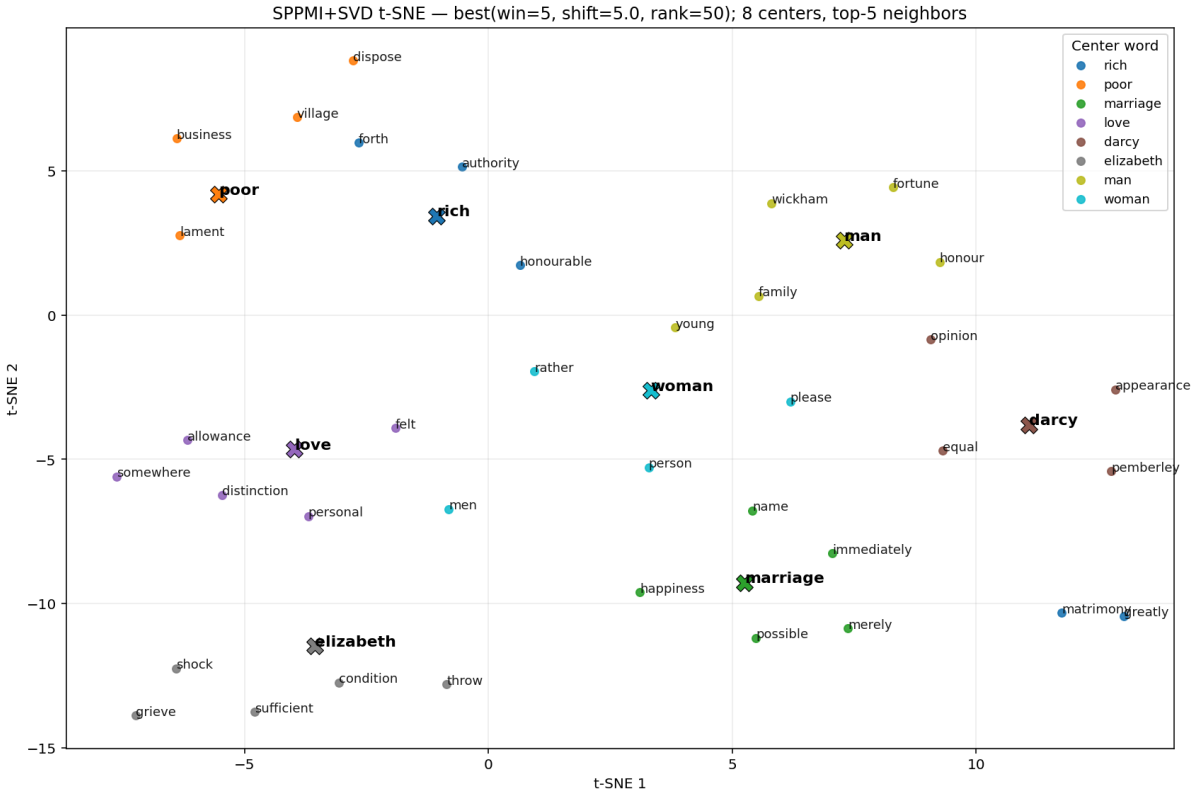
style axis, lack of phrase modeling ("mr_darcy"), and POS mixing.



4.2.3 SPPMI-SVD

The learned neighborhoods reveal systematic sociolinguistic patterns consistent with an Austen-era register. Gender terms exhibit asymmetry: man attracts status- and resource-laden associates (e.g., fortune, honour, Wickham, family), whereas woman co-occurs with relational or evaluative language (e.g., men, person, please, rather), indicating a patriarchal framing of identity through social ties and decorum. Class markers separate along governance versus precarity: rich aligns with institutional or moralized lexis (authority, honourable, matrimony), while poor draws terms of administration and hardship (business, dispose, village, lament), suggesting material constraint and social management. For character traits/actions, Elizabeth clusters with affect and stance (coolly, displease, shock, grieve), capturing a calibrated emotional register; Darcy centers on appraisal and standing (opinion, appearance, equal, Pemberley), reflecting judgment and estate identity. The marriage theme is framed procedurally and evaluatively—neighbors such as eligible, match, immediately, merely, name emphasize suitability, urgency, and social inscription rather than romantic idealization. We note that adverbials/particles (e.g., immediately, merely) and high-frequency stylistic axes may inflate some links; accordingly, interpretations rely on within-model comparisons and would benefit from phrase modeling and

POS-constrained analyses.



4.3 Comparison across models

Across all three models a consistent social grammar emerges, but with distinct emphases. Gender asymmetry is stable: men attract status/epistemic or resource-linked terms (e.g., fortune, honour, sensible/know), whereas women skew relational/affective and politeness markers (e.g., men/person/please, think/really), indicating a patriarchal framing of identity; GloVe amplifies dialogue and address tokens (e.g., mr, say), while SPPMI-SVD foregrounds evaluative stance around Elizabeth and status/estate identity around Darcy. Class signals are clearest in SPPMI-SVD (rich → institutional/moralized lexis; poor → administration/hardship), more diluted in Skip-gram (e.g., broad undeserving attaching to rich/man/woman), and mixed in GloVe (idiosyncratic associates such as qualification/stone). Marriage is procedurally framed in all models (e.g., eligible, match/offer, dispose, immediately), emphasizing suitability and social inscription over romance. Divergences often trace to modeling bias: GloVe's nearest neighbors are influenced by dialogue scaffolding, Skip-gram by corpus-wide sentiment/style axes, and both benefit from phrase modeling (e.g., mr_darcy) and POS constraints; thus we read cosine structure within models and treat cross-model differences as complementary evidence of the same underlying sociolinguistic script.

5. Conclusion

In this study, we evaluated and compared three word-embedding techniques—Skip-gram (Word2Vec), GloVe, and SPPMI–SVD—under a uniform pipeline on *Pride and Prejudice*. The findings provide actionable insight into how each method captures semantic relations in a small literary corpus: Skip-gram and GloVe were

competitively strong on our intrinsic neighbourhood metric, while SPPMI–SVD most clearly surfaced class/institutional vocabulary. All methods improved with larger vocabularies and more data, suggesting further gains from broader corpora or large pre-trained vectors. Our scope is limited to one corpus and an intrinsic evaluation; future work should include extrinsic tasks and additional embedding families. Overall, these results guide practitioners in selecting embeddings based on data size, interpretability, and domain needs.