Deriving semantic classes of Italian adjectives via word embeddings: a large-scale investigation

Ivan Lacić
University of Bologna
ivan.lacic2@unibo.it

Ludovica Pannitto
University of Bologna
ludovica.pannitto@unibo.it

Abstract

This paper investigates the application of word embeddings to derive semantic classes for Italian adjectives. Adjectives were clustered using UMAP for dimensionality reduction and K-means for clustering. Semantic categories such as "Relational", "Descriptive", "Evaluative", "Membership", and "Physical/Health-Related" were tested by employing predefined prototypical adjectives for each class. The precision and recall of the classification were analyzed, revealing high accuracy for some classes (e.g., "Evaluative"), but challenges in distinguishing more nuanced categories such as "Descriptive". Furthermore, cluster overlaps were visualized using KDE and quantified using KNN, , highlighting semantic intermingling between groups, especially between the "Descriptive" and "Evaluative" categories. Finally, a comparison with Wordnet's adjective categories was provided.

1 Introduction

Meaning is a fundamental aspect of language, making semantics essential for all levels of linguistic analysis. However, incorporating semantics into such analysis presents challenges due to the complexity and labor-intensive nature of semantic annotation. It is widely acknowledged that, unlike nouns and verbs, adjectives exhibit non-trivial semantic behavior, resulting from the intricate interaction between their semantic and syntactic properties. The meaning of adjectives is particularly fluid, often shifting based on linguistic context. Consider, for example, the adjective *heavy*, that can refer to physical weight in the sentence "The box is heavy", but takes on a different meaning in "It has been a heavy week", where it signifies emotional or mental strain rather than physical weight. As a result, analyzing and representing the semantics of adjectives is far from straightforward. WordNet (Miller, 1995) traditionally lacks a comprehensive semantic hierarchy for adjectival meanings, offering only a coarse classification with three labels derived from `lexfiles`: *adj.all* for descriptive adjectives, *adj.pert* for pertainyms, and *adj.ppl* for adjectival participles. Given the significant influence of linguistic context on adjectival meaning, this study explores the possibility of deriving a semantic classification of Italian adjectives through word embeddings, leveraging distributional semantics (Lenci and Sahlgren, 2023). By providing an empirical framework for categorizing adjectives based on their semantic similarities, this analysis highlights the advantages of using word embeddings for semantic classification while also identifying the limitations of current clustering techniques when applied to highly polysemous word classes.

The paper is structured as follows. Section 2 provides a concise overview of the current state of the art. Section 3 details the dataset used to construct the vector space. In Section 4, three case studies are introduced, along with their respective results. Section 5 compares the semantic classes derived from word embeddings with those in WordNet's classification. Lastly, Section 6 summarizes the key findings and offers recommendations for future research.

2 Related work

To the best of our knowledge, the only WordNet-derived resources that organize adjectival synsets into a hierarchy are GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010), which divides adjectives into 16 semantic classes based on those proposed by Hundsnurscher and Splett (1982), and the Bulgarian WordNet (Dimitrova and Stefanova, 2018), which largely follows the German approach. In addition to WordNet, several other feature-based semantic classification systems are available (Schweinberger and Luo, 2024). Typology-based approaches, such as those proposed by Dixon

(1977), offer language-independent classifications of adjectives based on their syntactic and morphological properties. Corpus-based classification systems, like the one in Biber et al. (2007), categorize adjectives based on frequency patterns in the Longman Spoken and Written English Corpus. Additionally, automated classification, such as the UCREL semantic analysis system (USAS, Piao et al. 2005), assigns words and MWEs to 21 semantic fields using resources such as lexical databases and thesauri. Although the aforementioned approaches provide valuable insights, they often rely on predefined structures or patterns that may not fully capture the semantic nuances of adjectives, particularly given their contextual variability. Furthermore, while some of these resources have been adapted for Italian (e.g., Python Multilingual UCREL Semantic Analysis System[1] , Songlin Piao et al. 2016), a comprehensive, Italian-specific semantic classification is still lacking.

Distributional approaches have been extensively applied to adjectival meaning, for example, to derive adjectival scales (Kim and de Marneffe, 2013) or their negated meaning (Aina et al., 2019). However, more focused works on the semantic classification of adjectives are scarcer. One such example is Montes and Geeraerts (2022), which explore the application of distributional methods to the semantic analysis of Dutch adjectives. The work also addresses the difficulties in aligning distributionally derived senses with lexicographic inventories, emphasizing how distributional models offer empirically founded categories that can be quantitatively described, explained in terms of contextual elements, and interpreted in terms of, at least, *aspects* of senses. Additionally, given their syntactic functions, Baroni and Zamparelli (2010) distributionally represent adjectives as matrices rather than vectors, interpreting them as linear maps that can be applied over nouns to shift their meaning. However, none of the above studies specifically focus on Italian adjectives.

3 Data

In order to test our intuition, we built a Distributional Space Model using the itWaC corpus (Baroni et al., 2009), a fairly large corpus (2 billion tokens) constructed by crawling the web from medium-frequency words from the Repubblica corpus (Baroni et al., 2004) and basic Italian vocabulary lists as seeds. The corpus comes already lemmatized and POS-tagged, which allowed us to directly isolate lexemes for representation in the distributional space, without requiring additional preprocessing steps. We utilized the `gensim`[2] (Řehřek and Sojka, 2010) implementation of the `word2vec` algorithm (Mikolov, 2013) to build the embeddings.More specifically, after evaluating model performance on Multilingual SimLex-999 and WS-353 (Vulić et al., 2020), we opted for the Skip-gram architecture over a 5-dimensional window to build 500-dimension vectors (see Appendix A). Out of a total of $12,812$ lemmas tagged as `ADJ` in the corpus, we filtered down to $8,348$ types by applying several filtering steps. We excluded adjectives with a frequency lower than 10 and removed non-existent and non-Italian adjectives, identified using criteria such as word length and cross-referencing the dataset with the `kaikki.org` machine-readable Italian dictionary (Ylonen, 2022). While this filtering approach may have result in the exclusion of some relevant adjectives, the benefits of this process were deemed to outweigh the potential drawbacks. This filtered space formed the basis for the subsequent clustering steps.

4 Experiments and results

This chapter presents the key analyses and findings of the study. We begin by investigating the construction of meaning clusters using dimensionality reduction techniques, as outlined in Section 4.1. Next, we evaluate the semantics of these clusters by comparing them with predefined categories of adjectives, as detailed in Section 4.2. Finally, Section 4.3 examines the overlap between clusters to assess the degree of semantic intermingling between groups.

4.1 Looking for meaning clusters

The first step of the analysis involved exploring the constructed distributional space by reducing the high-dimensional vector space to a two-dimensional plane using Uniform Manifold Approximation and Projection (UMAP, Leland et al. 2018) with default parameters (*n-neighbors* $= 15$, *min-dist* $= 0.1$). UMAP, like the widely employed t-SNE algorithm, generates a high-dimensional graph representation of the data and optimizes a low-dimensional graph to closely match the

[1] `https://github.com/UCREL/pymusas?tab=readme-ov-file`

[2] `https://radimrehurek.com/gensim/`

original structure. However, compared to t-SNE, UMAP is generally more effective at preserving the global structure in the final projection. Figure 1 displays the UMAP projection, where denser areas indicate groups of adjectives with similar distributional patterns, while more isolated points represent adjectives with unique distributional characteristics.

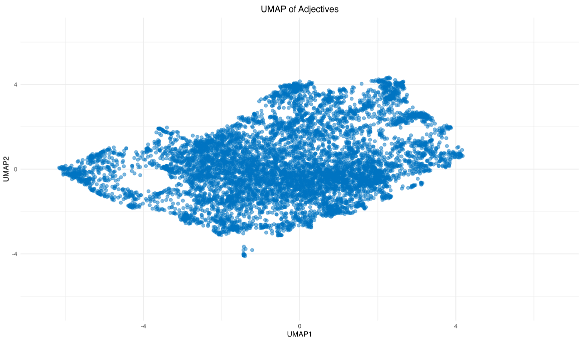

Figure 1: UMAP dimensionality reduction

To determine the optimal number of clusters, i.e. groups, into which the dataset should be divided, the `NbClust` package (Charrad et al., 2014) was used. `NbClust` computes up to 30 indices to determine the appropriate number of clusters and recommends the best clustering solution based on a majority vote among these indices. For this study, the index argument was set to "all", resulting in the calculation of 26 indices (excluding *Gamma*, *Tau*, *Gap*, and *Gplus* due to their high computational costs). Based on 10 indices[3], including the D-index (Figure 2), it was shown that 5 clusters represent the optimal clustering solution. The D-index (Lebart et al., 1995) showed a marked improvement up to 5 clusters, with diminishing returns beyond that point. The D-index is based on clustering gain on intra-cluster inertia, namely the degree of homogeneity among data pertaining to a cluster. It is computed for each step P_k, consisting of k clusters, as the average distance between each point assigned a cluster and the cluster centroid. Given two partitions P_{k-1} and P_k, composed of $k-1$ and k clusters respectively, the gain in intra-cluster inertia is defined as (with $d(P_k)$ being the value of the index at k clusters):

$$G = d(P_{k-1}) - d(P_k) \tag{1}$$

The optimal number of clusters can, therefore, be visually identified by the sharp knee in the graph, which corresponds to a significant decrease in inertia gain. The noticeable drop in D-index (Figure 2 – left plot) and a large spike in the second differences plot (Figure 2 – right plot) suggest the point of the most significant improvement in clustering.

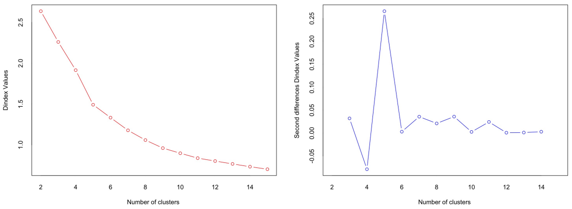

Figure 2: D-index: the left side of the panel shows the D-index itself while the right side shows the gain in intra-cluster inertia.

Consequently, the UMAP-reduced data was clustered into 5 groups using *K-means*, as shown in Figure 3. K-Means partitions a set of n observations into K clusters, with each observation assigned to the cluster whose centroid is nearest, serving as the representative instance of that cluster. Table 1 presents the distribution of the $8,348$ adjectives across the clusters.

Cluster	# of adjectives	Label
1	467	Relational
2	2436	Descriptive
3	2709	Evaluative
4	1038	Physical/Health related
5	1698	Membership

Table 1: Distribution of adjectives in five clusters. Labels in the third column are discussed in Section 4.2

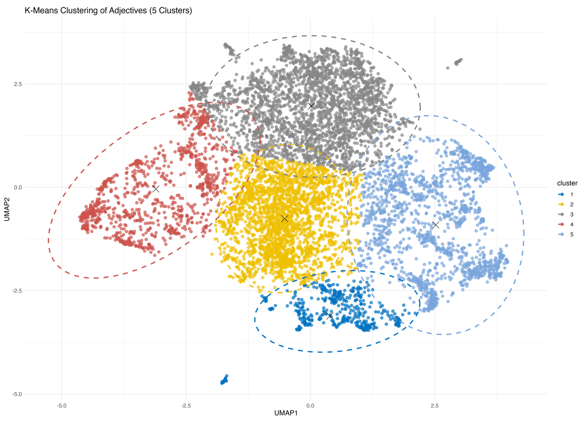

Figure 3: K-Means clustering: in the plot, each dot represents and adjective colored according to its cluster

The quality of the clustering was evaluated using the Davies-Bouldin Index (Davies and Bouldin, 1979) and the Calinski-Harabasz Index (Caliński

[3] KL, CCC, Scott, TraceW, Friedman, Rubin, DB, Ratkowsky, Dunn, SDindex.

and Harabasz, 1974), with scores of 0.92 and 7119.29, respectively, indicating satisfactory and distinct clusters.

4.2 Evaluation of the semantics of clusters

We now turn to examining the adjectives within these clusters to establish five labels that represent semantic categories for the classification of adjectives

The initial approach involved extracting the top 50 most representative adjectives for each cluster, based on their Euclidean distance to the cluster centroid. This analysis revealed that Cluster 1 predominantly includes environmental and agricultural adjectives (e.g., *meteoclimatico* 'meteoroclimatic', *orticolo* 'horticultural') adjectives, while Cluster 2 mainly features adjectives expressing temporal and relational relationships (e.g., *indiretto* 'indirect', *futuro* 'future'). Cluster 3 is characterized by emotional and evaluative adjectives (e.g., *indemoniato* 'possessed', *soffocante* 'suffocating'). In Cluster 4, technical adjectives, mainly related to chemistry, are prevalent (e.g., *chimico* 'chemical', *inorganico* 'inorganic'). Finally, Cluster 5 is centered around historical and cultural adjectives, such as *aristocratico* 'aristocratic' and *feudale* 'feudal'. However, such approach was deemed overly fine-grained and influenced by the specificity of the 50 adjectives extracted. Consequently, a larger random sample of adjectives from each cluster was manually examined to refine the classification. Upon careful analysis, the refined labels presented in Table 2 were proposed, along with examples of adjectives for each label.

Cluster	Adjectives	Label
1	*primo* 'first', *ambientale* 'environmental', *idrico* 'aquatic'	Relational (Rel)
2	*nuovo* 'new', *specifico* 'specific', *rotondo* 'round'	Descriptive (Des)
3	*bello* 'nice', *difficile* 'hard', *eccessivo* 'excessive'	Evaluative (Eva)
4	*sanitario* 'healthcare', *cronico* 'chronic', *chimico* 'chemical'	Physical/Health-Related (Phy/H)
5	*italiano* 'Italian', *musulmano* 'Muslim', *democratico* 'democratic'	Membership (Mem)

Table 2: Adjective clusters and suggested labels

Recognizing the contextual flexibility of adjectives, it is evident that certain adjectives may be associated with multiple labels, as the proposed labels are not rigid but interconnected. For instance, *Membership* and *Physical/Health-Related* classes can be considered sub-classes of *Relational adjectives*, as they share inherent relational traits. However, distinct clusters in the visualization highlight unique contextual patterns that differentiate these groups. Additionally, while the suggested labels reflect the predominant themes of each cluster, it is important to note that not every adjective within a cluster will strictly conform to the assigned category. This is because *K-means* clustering, applied to general-purpose word embeddings, captures patterns of co-occurrence and contextual similarity rather than rigid or strictly contextual semantics. To validate the consistency of the semantic categories, each adjective was compared with a set of representative examples for each identified semantic category, assigning the adjective to the most pertinent label. We manually identified six highly prototypical adjectives (by asking fellow expert linguists) for each target category ($p_i^1, ...p_i^6$ for clusters $C_i \in [1, 5]$) based on the five aforementioned semantic classes. These six adjectives per category are presented in Table 3.

Cluster	Adjectives
Rel	*primo* 'first', *ultimo* 'last', *notturno* 'nocturnal', *architettonico* 'architectural', *costiero*, 'coastal', *idrico* 'water-related'
Des	*nuovo* 'new', *vecchio* 'old', *rotondo* 'round', *silenzioso* 'quiet', *grande* 'big', *verde* 'green'
Eva	*bello* 'nice', *buono* 'good', *orrendo* 'horrible', *cattivo* 'mean', *stupido* 'stupid', *fantastico* 'fantastic'
Phy/H	*chimico* 'chemical', *biologico* 'biological', *ospedaliero* 'hospital-related', *genetico* 'genetic', *visivo* 'visual', *malato* 'sick'
Mem	*italiano* 'Italian', *americano* 'American', *democratico* 'democratic', *straniero* 'foreign', *domestico* 'domestic', *cristiano* 'Christian'

Table 3: Six prototypical adjectives per category

From each of the 5 clusters C_i, 250 adjectives $a_i^1, ...a_i^{250}$ were selected to assess their alignment with predefined semantic categories. For each adjective a_i^n, we retrieved its 30 nearest neighbors $N(a_i^n, 30)$ in the vector space and calculated the cosine similarity between these neighbors and each of the six prototypical adjectives $p_j^{\{1,...,6\}}$. Each adjective a_k^n was then assigned to the semantic class C_s that maximized the cumulative similarity between its 30 nearest neighbors $N(a_k^n)$ and the prototypical adjectives $p_s^1, ..., p_s^6$ in each semantic class.

In formula, this can be expressed as:

$$C(a_k^n) = \arg\max_s \left(\sum_{x \in N(a_k^n)} \sum_{p \in \{p_s^1, \ldots p_s^6\}} \cos(x, p) \right) \quad (2)$$

This scoring system aims to match each adjective with the semantic category that most accurately reflects its typical usage based on vector space relationships. After identifying the most probable category for each adjective, we evaluated the overall alignment of each cluster with the semantic category represented by the prototype adjectives. This evaluation focused on how well the clustering-based labels aligned with the semantic themes identified by human informants. The results of the validation are summarized in Table 4. Categories such as *Evaluative* and *Relational* are predominantly represented by adjectives that align with their initial cluster label. Others, like *Descriptive*, exhibit a broader distribution of adjectives across multiple clusters.

	Rel	**Des**	**Eva**	**Phy/H**	**Mem**	TOT
Rel	**189**	8	7	37	9	250
Des	56	**69**	51	28	46	250
Eva	5	20	**216**	0	9	250
Phy/H	10	23	30	**179**	8	250
Mem	23	13	19	21	**174**	250
TOT	283	133	323	265	246	

Table 4: Confusion matrix presenting the validation results

To quantify the agreement between the cluster-assigned labels and the prototype-derived semantic categories, Precision, Recall, and F1-Score were calculated (Table 5). As observed, precision varies considerably across the clusters. The *Evaluative* cluster achieves the highest precision (0.864), indicating accurate classification with few false positives. In contrast, the *Descriptive* cluster exhibits the lowest precision (0.276), reflecting the challenges in distinguishing descriptive adjectives from other classes. Recall, on the other hand, is moderate to high for most clusters, with the *Membership* cluster performing the best (0.702).

Overall, the metrics suggest that the clustering strategy is effective, while also revealing challenges in distinguishing overlapping or subtly distinct categories. This is reflected in a Normalized Mutual Information (NMI) score of 0.3724 and an Adjusted Rand Index (ARI) of 0.3756, suggesting moderate agreement between the prototype-derived categories and the cluster-derived labels. Additionally, the similarity between the two metrics suggests that the clusters are not significantly affected by the chance agreement, thereby strengthening the overall validity of the clustering results. The relatively lower NMI and ARI scores may be attributed to semantic overlaps (adjectives may belong to multiple categories), cluster ambiguity (clusters contain mixed or overlapping semantic categories), or the limitations of the *K-means* algorithm when applied to complex linguistic data.

4.3 Cluster overlap

To further investigate one potential cause, viz. overlap between clusters, Kernel Density Estimation (KDE, Silverman 1998) was employed to visualize regions of high density within the clusters. KDE generates a continuous density surface, highlighting areas where adjectives are densely packed and where clusters overlap. Dense regions represent the core areas of each cluster, most indicative of each cluster's semantic space, while overlapping contours indicate shared or closely positioned adjectives. The resulting density plot is shown in Figure 4.

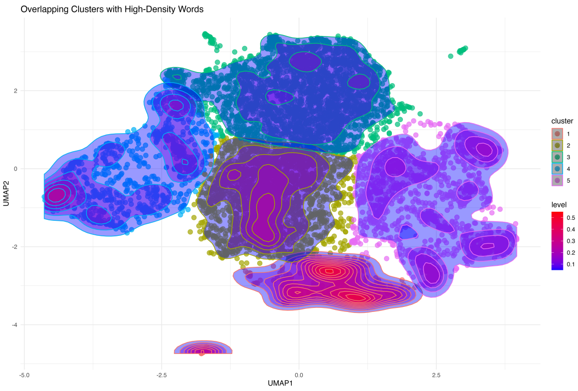

Figure 4: Kernel Density Estimation (KDE) plot

The high-density contours in the plot demonstrate that the clusters generally maintain their distinct boundaries, with little overlap between them. While clusters may touch or be adjacent, their core high-density areas remain separate, indicating clear semantic distinctions. To further investigate the overlap between clusters, we proceed by quantifying it, using spatial indexing with a *K-nearest neighbor* (KNN) approach[4], provided by the `Fast`

[4] A combined approach using KNN and Gaussian Mixture Models (GMM) with Bhattacharyya distance was tested to quantify overlap based on probability distributions, incorporating weighted averages. However, it was determined that the KNN values alone provided more interpretable and straightforward results

Cluster	TP	FP	FN	TN	Precision	Recall	F1-Score
Relational	189	61	88	912	0.756	0.682	0.717
Descriptive	69	181	63	937	0.276	0.522	0.361
Evaluative	216	34	110	890	0.864	0.663	0.750
Physical/Health	179	71	88	912	0.716	0.670	0.692
Membership	174	76	74	926	0.696	0.702	0.699

Table 5: Classification metrics for each cluster

`Nearest Neighbor` (FNN) package[5]. The overlap between two clusters was operationalized as the number of points in the first cluster that have a nearby point in the second cluster (within a distance threshold 0.5 units in the UMAP space). The overlap counts are presented in Table 6.

	Rel	Des	Eva	Phy/H	Mem
Rel	x	x	x	x	x
Des	131	x	x	x	x
Eva	0	399	x	x	x
Phy/H	0	233	209	x	x
Mem	127	144	113	0	x

Table 6: Overlaps between clusters – K-Nearest Neighbors

The greatest overlap was observed between the *Descriptive* and *Evaluative* cluster (399 points). These findings corroborate the classification results from the 250 adjective samples (see Table 4), where a substantial portion of the adjectives intended for the *Descriptive* cluster was frequently misclassified as *Evaluative*. The substantial number of shared points between these two clusters emphasizes that adjectives assigned to the *Evaluative* and *Descriptive* categories are not only close in semantic space but are often intermingled, making it difficult to separate them in a reduced-dimensional representation. The overlap detected using the KNN approach thus serves as a validation of the classification confusion observed, confirming that the semantic boundaries between *Descriptive* and *Evaluative* adjectives are porous, leading to a blending of meanings that complicates their discrete categorization.

Following the cluster evaluation, we put the annotation scheme to the test and proceed with the annotation of a sample of 500 adjectival lexemes extracted from itWaC. Unlike the highly prototypical adjectives used in the previous tests, this sample was not controlled, containing both frequent, unambiguous adjectives and those with lower frequencies and less straightforward semantics. The labeling results are presented in Table 7.

Label	TOT
Relational	7 (1.40%)
Descriptive	143 (28.60%)
Evaluative	285 (57.00%)
Physical/Health	31 (6.20%)
Membership	34 (6.80%)

Table 7: Results of word-embedding based semantic annotation

It can be observed that the majority of adjective types were classified as *Evaluative* and *Descriptive* adjectives, while only seven types were categorized as *Relational* adjectives. Following the initial automated clustering, a manual review was conducted to assess the accuracy of the labels assigned to the adjectives during the annotation process. Although the overall classification was generally satisfactory, the review revealed some instances of misclassification. In certain cases, it was determined that an alternative classification might more accurately reflect the semantic nature of the adjectives. Given that both the word embeddings and the chosen clustering method reflect patterns of co-occurrence and contextual similarity rather than strict semantic similarity, such results are not entirely unexpected. For instance, the assignment of the Membership class to the adjective *criminale* ‘criminal’ can be explained by the fact that the most strongly associated nouns (according to log-likelihood ratio) of the adjective in question in the itWaC corpus – *organizzazione* ‘organization’, *banda* ‘gang’, and *gruppo* ‘group’ – are all closely tied to concepts of membership and community. Despite these occasional discrepancies, we argue that the classification has provided a solid foundation and yielded valuable insights.

5 Wordnet comparison

Finally, we compared the proposed classification with the labels available in Wordnet, namely *adj.all* and *adj.pert*, as our dataset did not include any adjectival participles, categorized as verbs in itWaC.

By examining data in OpenMultilingualWordnet

[5] `https://cran.r-project.org/web/packages/FNN/index.html`

# Synsets Cluster	0	1	2	3	4	5	6	8	TOT
Relational	167	92	12	2	3	/	/	1	277
Descriptive	53	50	13	9	7	1	/	/	133
Evaluative	83	152	64	13	6	4	2	1	325
Physical/Health	116	131	16	2	1	/	/	/	267
Membership	89	134	20	3	/	/	/	/	246
TOT	508	559	125	30	16	6	3	1	1250

Table 8: Distribution of the number of synsets across clusters

(`omw-1.4`, Bond and Paik 2012; Bond et al. 2016), we annotated a sample of 250 adjectives randomly extracted from each cluster. As far as Italian is concerned, data in `omw` is derived from MultiWordNet (Pianta et al., 2002), a version of the Italian WordNet aligned with Princeton WordNet 1.6. Coverage was incomplete, as 507 out of 1250 adjectives were not associated with any synset in the database. Of the remaining adjectives, 559 were linked to a unique synset, while the others resulted ambiguous (≥ 2 synsets) . Table 8 displays the distribution of the number of synsets associated with each adjective across the five clusters.

For each retrieved synset, we used the `nltk`[6] to extract the semantic label (i.e., *all* or *pert*), which, in the case of adjectives, corresponds to the `lexname`.

First, a strong correlation was observed between Wordnet's *pert* label and our relational macro-category, which encompasses the Relational, Physical/Health-Related, and Membership classes. Specifically, of the 272 types labeled as *pert*, 258 (94.85%) fell within this macro-category (46 Relational, 102 Physical/Health-Related, and 110 Membership), highlighting both the embeddings' ability to capture relational meanings and the interconnectedness of these three semantic labels. Second, for the *all* label, another interesting pattern emerged. Among 149 adjective types with more than one synset, 116 (77.85%) were found in the Descriptive and Evaluative classes, reinforcing the higher polysemy in these clusters. This aligns with the earlier analysis, which showed that these two clusters exhibited the greatest overlap.

6 Conclusion and future work

The paper explored the possibility of deriving semantic classes for Italian adjectives using word embeddings. While adjectives belonging to certain categories were easily identifiable and demonstrated high accuracy (e.g., the *Evaluative* class), others (e.g., the *Descriptive* class) proved more difficult to categorize and lacked consistency. Furthermore, it was observed that, as shown in previous literature, the collocational context of adjectives, particularly with regard to the nouns they modify, has a significant influence on their classification in semantic classes. A holistic approach, which considers the bidirectional relationship between adjectives and nouns, is hence essential (for the so-called *composition method* to account for polysemy effects, see, e.g., Baroni and Zamparelli 2010, who treat adjectives as data-induced (linear) functions over nominal vectors).

To sum up, the semantic classification of Italian adjectives using word embeddings is feasible but not without challenges. Besides the issues with cluster coherence, adjective polysemy must also be addressed, as semantic class categorization may also depend on the specific sense of an adjective (cf., for instance, *raffreddato* which, depending on the context, can mean both 'cooled/chilled', hence belong to a *Descriptive* class, as well as 'have a cold', thus belong to *Physical/Health related* class). In this regard, the choice of embeddings could impact the results. Although POS-tagging enables our static model to differentiate between senses with different parts of speech, senses that share the same POS are still merged into a single vector. Therefore, it would be valuable to explore vectors based on syntactic collocates rather than a linear window, such as those produced by `word2vecf` (Levy and Goldberg, 2014), or natively contextual embeddings like those generated by BERT (Devlin et al., 2019), as they might perform better than static, general-purpose ones (see, *inter alia*, Soper and Koenig 2022 for a discussion on this topic). In addition, tests with alternative classifiers should be run. Finally, it would be beneficial to explore different (higher) values for K, as the five clusters used in this study represent a relatively coarse-grained grouping. For studies requiring more precision (fine-grained classification), this may not be suffi-

[6] https://www.nltk.org/howto/wordnet.html

cient. Regarding WordNet, while the nominal side of WordNet is deeply structured (with a lot of levels), in this proposal we suggest adding just one additional level to the hierarchy.

Acknowledgments

The authors are grateful to Dr. Matteo Pascoli for his support in the creation of the vector spaces.

References

Laura Aina, Raffaella Bernardi, and Raquel Fernández. 2019. Negated adjectives and antonyms in distributional semantics: not similar? *IJCoL. Italian Journal of Computational Linguistics*, 5(5-1):57–71.

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, Marco Mazzoleni, et al. 2004. Introducing the La Repubblica corpus: A large, annotated, tei (xml)-compliant corpus of newspaper Italian. In *LREC*.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43:209–226.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 2007. *Grammar of spoken and written English*. Longman.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th global WordNet conference (GWC 2012)*, pages 64–71. Matsue.

Francis Bond, Piek Vossen, John Philip McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 50–57.

Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. Nbclust: an R package for determining the relevant number of clusters in a data set. *Journal of statistical software*, 61:1–36.

David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 7–12.

Tsvetana Dimitrova and Valentina Stefanova. 2018. The semantic classification of adjectives in the Bulgarian wordnet: Towards a multiclass approach. *Cognitive Studies*, (18).

Robert MW Dixon. 1977. Where have all the adjectives gone? *Studies in Language*, 1:19–80.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT - The GermaNet Editing Tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2228–2235, Valletta, Malta.

Franz Hundsnurscher and Jochen Splett. 1982. *Semantik der Adjektive des Deutschen: Analyse der semantischen Relationen*. Springer.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1625–1630.

Ludovic Lebart, Alain Morineau, and Marie Piron. 1995. *Statistique exploratoire multidimensionnelle*. Dunod.

McInnes Leland, Healy John, and Melville James. 2018. Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Alessandro Lenci and Magnus Sahlgren. 2023. *Distributional semantics*. Cambridge University Press.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mariana Montes and Dirk Geeraerts. 2022. How vector space models disambiguate adjectives: A perilous but valid enterprise. *Yearbook of the German Cognitive Linguistics Association*, 10(1):7–32.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Scott Songlin Piao, Paul Rayson, Dawn Archer, and Tony McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech Language*, 19(4):378–397. Special issue on Multiword Expression.

Radim Řehřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. University of Malta.

Martin Schweinberger and Chang-Hao (Howard) Luo. 2024. Automated, corpus- and usage-based semantic classification of word class using word embeddings. Presentation delivered at ICAME 45 "Interlocking corpora and register(s): Diversity and innovation"'. Vigo, 18–22 June 2024.

Bernard W Silverman. 1998. *Density estimation for statistics and data analysis*. Routledge.

Scott Songlin Piao, Paul Edward Rayson, Dawn Archer, Francesca Bianchi, Carmen Dayrell, Mahmoud El-Haj, Ricardo-María Jiménez-Yáñez, Dawn Knight, Michal Křen, Laura Lofberg, et al. 2016. Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. In *LREC 2016, Tenth International Conference on Language Resources and Evaluation*.

Elizabeth Soper and Jean-Pierre Koenig. 2022. When polysemy matters: Modeling semantic categorization with word embeddings. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 123–131.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, et al. 2020. Multi-simlex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC)*, pages 1317–1325.

A Distributional models

Models were created both with `CBoW` and `SkipGram` algorithms, with (linear) windows of 2 and 5 tokens and embedding dimensions of 200, 300, 500. Table 9 summarizes the results of the correlation with the popular dataset of Word Similarity and Relatedness.

Algorithm	**CBOW**						**SkipGram**					
Window	**win2**			**win5**			**win2**			**win5**		
Dimension	**200**	**300**	**500**	**200**	**300**	**500**	**200**	**300**	**500**	**200**	**300**	**500**
SimLex-999	0,343	0,353	0,370	0,342	0,349	0,364	0,354	0,373	**0,398**	0,334	0,352	0,370
WS353	0,519	0,524	0,519	0,558	0,554	0,555	0,559	0,562	0,565	0,569	0,571	**0,576**
WS353-Rel	0,402	0,408	0,396	0,455	0,447	0,444	0,466	0,473	0,459	0,486	**0,487**	**0,487**

Table 9: Correlation of cosine similarity with scores in Italian SimLex-999 and WS-353. Coverage was 924 out of 999 items for SimLex, and 287 out of 350 items for WS-353.