Information Processing and Management 59 (2022) 102925

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

# Learning interpretable word embeddings via bidirectional alignment of dimensions with semantic concepts

Lütfi Kerem Şenel [a,*], Furkan Şahinuç [b,c,d], Veysel Yücesoy [c], Hinrich Schütze [a], Tolga Çukur [b,d], Aykut Koç [b,d]

[a] Center for Information and Language Processing (CIS), LMU Munich, Germany
[b] Electrical and Electronics Engineering Department, Bilkent University, Ankara, Turkey
[c] ASELSAN Research Center, Ankara, Turkey
[d] National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara, Turkey

ARTICLE INFO

*Keywords:*
Word embeddings
Interpretability
Word semantics

ABSTRACT

We propose *bidirectional imparting* or *BiImp*, a generalized method for aligning embedding dimensions with concepts during the embedding learning phase. While preserving the semantic structure of the embedding space, BiImp makes dimensions interpretable, which has a critical role in deciphering the black-box behavior of word embeddings. BiImp separately utilizes both directions of a vector space dimension: each direction can be assigned to a different concept. This increases the number of concepts that can be represented in the embedding space. Our experimental results demonstrate the interpretability of BiImp embeddings without making compromises on the semantic task performance. We also use BiImp to reduce gender bias in word embeddings by encoding gender-opposite concepts (e.g., male–female) in a single embedding dimension. These results highlight the potential of BiImp in reducing biases and stereotypes present in word embeddings. Furthermore, task or domain-specific interpretable word embeddings can be obtained by adjusting the corresponding word groups in embedding dimensions according to task or domain. As a result, BiImp offers wide liberty in studying word embeddings without any further effort.

## 1. Introduction

Developments in machine learning lead to interdisciplinary studies and merge different research areas. An example can be observed in the natural language processing (NLP) based information science studies. There are increasingly improving information science studies that utilize NLP methods, especially word embeddings, while focusing on processing textual information. The scope of NLP-based studies can range from event detection (Qian et al., 2019; Tuke et al., 2020) to document retrieval (Bagheri et al., 2018). Computational studies on social media also frequently utilize NLP tools in various topics such as author profiling (López-Santillan et al., 2020), content processing (Moudjari et al., 2021; Roy et al., 2021) and hate speech detection (Pamungkas et al., 2021; Pronoza et al., 2021). What is common among these studies is that they all heavily depend on textual data. In representing and processing text, word embeddings play a key role and are used ubiquitously. Word embeddings are pre-trained semantic representations of words that hold numerous semantic features of natural languages. However, one disadvantage of word embeddings is that they learn language features as black-box schemes, unlike methods directly extracting determined and desired features. Therefore, studies on

[*] Corresponding author.
*E-mail address:* lksenel@cis.lmu.de (L.K. Şenel).

https://doi.org/10.1016/j.ipm.2022.102925
Received 26 October 2021; Received in revised form 6 February 2022; Accepted 1 March 2022

their interpretability are of importance (Chen et al., 2016; Levy & Goldberg, 2014) for developing explainable NLP methods to be used in higher level information science applications.

*Word embeddings* (Bojanowski et al., 2017; Mikolov, Corrado et al., 2013; Mikolov, Sutskever et al., 2013; Pennington et al., 2014) – continuous dense vector representations – capture semantic and syntactic features of words. These embeddings are shown to be useful in a broad range of NLP applications involving topic modeling (Zhao et al., 2021), text classification (Elnagar et al., 2020), key-phrase extraction (Papagiannopoulou & Tsoumakas, 2018), document retrieval (Bagheri et al., 2018), named entity recognition (NER) (Nozza et al., 2021), query performance prediction (Roy et al., 2019), and extracting semantic features of words (Şahinuç & Koç, 2021). Although contextualized word embeddings and transformer-based architectures (Devlin et al., 2019; Radford et al., 2019; Vaswani et al., 2017) are becoming more and more prevalent due to their impressive performance on many NLP tasks, these models still use a static word embedding layer to represent input. Therefore, improvements to static word embeddings can potentially be transferred to contextual models as well (Schick & Schütze, 2020).

In addition to the traditional NLP tasks, word embeddings are frequently used in many other interdisciplinary domains. In neuroscience, they are employed to analyze the representation of semantics in brain activity (Huth et al., 2016; Ruan et al., 2016; Zhang et al., 2020) and as part of a decoder that extracts linguistic meaning from measured brain activity (Pereira et al., 2018). In psychiatry, they are used to detect incoherent speech for diagnosing schizophrenia (Iter et al., 2018; Voppel et al., 2021). In legal domain, they are used to predict outcomes of courts (Mumcuoğlu et al., 2021), evidence extraction from court records (Ji, Tao et al., 2020) and coreference resolution in legal texts (Ji, Gao et al., 2020). In the social domain, based on word, sentence and document embeddings polarization in social media can be analyzed (Demszky et al., 2019) and users of social media can be profiled (López-Santillan et al., 2020). Evolutionary linguists track historical changes in word meaning with embeddings (Hamilton et al., 2016; Kutuzov et al., 2018; Yüksel et al., 2021). Recent studies suggest that embeddings capture and quantify gender and ethnic biases in language (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018) and their evolution over time (Agarwal et al., 2019).

Despite a large body of work on improved word embeddings (Bollegala et al., 2016; Celikyilmaz et al., 2015; Liu et al., 2015; Mrkšić et al., 2016; Yang & Mao, 2016; Yu & Dredze, 2014; Yu et al., 2017), a central limitation is their lack of interpretability: dimensions of the dense vector space do not individually represent semantic concepts (Chen et al., 2016; Levy & Goldberg, 2014) or other directly interpretable distinctions. Yet interpretability of word embeddings is highly desirable for several reasons. (i) It will enable researchers to make sense of embeddings of individual words, which are currently meaningful only in relation to other embeddings. (ii) Word embeddings serve as base representation in many deep learning models, so their interpretability is key for interpretable deep learning models. (iii) In interpretable embedding models, it is easier to remove redundant or nonrelevant dimensions, resulting in reduced computation and memory requirements. (iv) Interpretability also facilitates removal of gender, race and other biases (Dufter & Schütze, 2019).

Previous studies have put forth several important approaches to address limitations on interpretability of word embeddings. A group of studies proposed to use sparsity constraints such as non-negative matrix factorization (Fyshe et al., 2014; Luo et al., 2015; Murphy et al., 2012), sparse coding (Arora et al., 2018; Faruqui et al., 2015) and sparse auto-encoders (Subramanian et al., 2018) that yield sparse word representations. Since each word is represented by only a few dimensions, it is easier to understand what semantic features the dimensions capture. However, larger vocabulary requires higher dimensionality to achieve a desired sparsity level which increases memory and computation requirements. In addition, evaluations on common benchmark tests suggest that the resulting sparse embeddings often perform poorly compared to the dense embeddings that have distributed word representations. Another group of studies proposed to instead use orthogonal transformations over the high performing dense embeddings (Dufter & Schütze, 2019; Park et al., 2017; Zobnin, 2017) in order to preserve task performance. Yet, the level of improvement in interpretability that orthogonal transformations can achieve is relatively limited. Recently, in Şenel et al. (2020), we proposed an offline imparting approach to obtain interpretable word embeddings by modifying the objective function of GloVe (Pennington et al., 2014) to align each dimension of the vector space with a single pre-defined concept. However, this unidirectional imparting method does not utilize the full capacity of the embedding space (negative directions are ignored) and is limited to the training setting of the GloVe.

In this paper, we introduce *BiImp* (read as "bimp"), a generalized imparting approach that is capable of bidirectional imparting and online learning, hence more efficient and adaptable to new training data. BiImp utilizes both directions along each dimension of the vector space separately to encode two different concepts. The two concepts can be chosen arbitrarily or chosen as opposites (e.g., *good – bad*, *male – female*) as a special case (see Fig. 1), providing a more efficient use of the embedding space while increasing encoding flexibility. We demonstrate BiImp by modifying the word2vec skip-gram model (Mikolov, Corrado et al., 2013; Mikolov, Sutskever et al., 2013); concepts are selected from Roget's Thesaurus and WordNet. A hyperparameter can be tuned to achieve a good tradeoff between interpretability on the one hand and preservation of semantic structure on the other. We perform comprehensive experiments and demonstrate that interpretability of word embeddings improves while performance stays about the same. Inspired by Bolukbasi et al. (2016), we also demonstrate that BiImp can concentrate gender information in a single embedding dimension, the gender dimension, as a continuum. This supports efficient capture of gender bias and debiasing through removal of the gender dimension. In short, main outcomes of this study can be summarized as: (i) BiImp provides interpretable word embeddings by using both positive and negative directions of word embeddings; (ii) BiImp is compatible to different word embedding learning types; (iii) BiImp can be utilized to remove human biases from embeddings without compromising task performance.

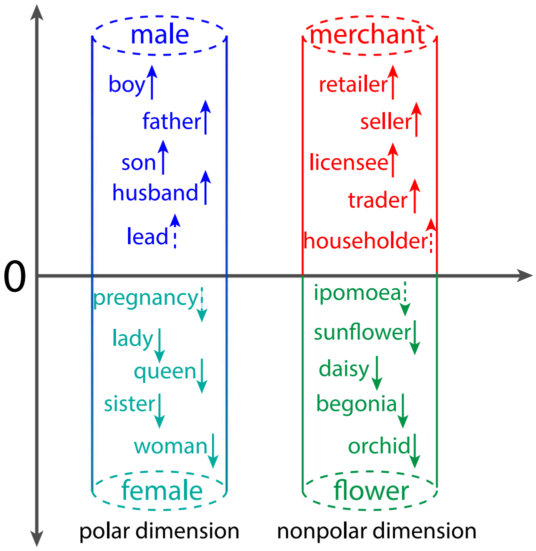

**Fig. 1.** Illustration of bidirectional imparting, the main idea underlying BiImp. The method increases interpretability of word embeddings by linking embedding dimensions to concepts. The concepts are taken from a conceptual resource that provides concepts along with word sets that are associated with them. BiImp "imparts" two concepts to each embedding dimension, one for the positive, one for the negative direction. E.g., the concepts "male" and "female" are associated with positive and negative directions of the polar dimension in the figure. Imparting is achieved by modifying the embedding training objective: during training, words associated with a concept are constrained to have high (or low) values on the dimension linked to the concept. As a result, the embedding vector of a word is directly interpretable: the value of each coordinate can be seen as a weight that the associated concept (positively or negatively associated concept) has in the representation of the meaning of the word. We study both polar dimensions (positive/negative concepts are opposites) and nonpolar dimensions (positive/negative concepts are unrelated). Solid arrow: word from resource. Dashed arrow: word not from the resource inferred to be related to the concept. We show that BiImp increases interpretability without impacting task performance and that it supports more effective debiasing.

## 2. Related work

### 2.1. *Interpretability of word embeddings*

Benefits of interpretable word embeddings have motivated several previous efforts to improve interpretability. Most of these studies introduce a sparsity constraint to learn sparse representations where each word is represented by only a few non-zero dimensions. The motivation behind sparsity is that by investigating the words that correspond to non-zero values in a dimension, one can infer which semantic features are encoded in that dimension. Based on this idea, Murphy et al. (2012) propose non-negative sparse embeddings (NNSE) to perform non-negative matrix factorization (NMF) on word co-occurrence variant matrices. As an extension to NNSE, Fyshe et al. (2014) proposed joint non-negative sparse embeddings (JNNSE) to incorporate additional knowledge on word similarity as measured by the similarity of cortical activity patterns. To address the memory and scale issues of NNSE-based methods, Luo et al. (2015) proposed an online learning method, where sparse embeddings were obtained using a modified skip-gram model (Mikolov, Sutskever et al., 2013). Several other studies proposed to learn sparse transformations that map pretrained state-of-the-art embeddings to sparse, more interpretable vector spaces instead of learning them from corpora (or co-occurrence matrices) directly. Arora et al. (2018) and Faruqui et al. (2015) use sparse coding methods and Subramanian et al. (2018) train a sparse auto-encoder. Inspired by research in topic modeling, Panigrahi et al. (2019) proposed a method named Word2Sense based on the Latent Dirichlet Allocation (LDA) to extract distributions of difference word senses from a corpus, which are then used to learn sparse interpretable word embeddings. While the above-mentioned approaches can increase interpretability to a certain degree, they do not exercise control over the specific concepts or word senses that are captured in the embedding dimensions.

Sparse representations typically have higher dimensionality than dense embeddings since only a few words are encoded in each dimension. Thus, they can suffer from memory and scaling issues especially for tasks that require a large vocabulary. To strictly preserve the dimensionality and semantic structure of word embeddings, several researchers proposed orthogonal instead of sparse transformations. Park et al. (2017) experimented with rotation algorithms based on exploratory factor analysis (EFA) with orthogonality constraints. Zobnin (2017) used orthogonal transformations to improve clustering of words along individual embedding dimensions. However, increases in clustering along a subset of embedding dimensions come at the expense of reduced clustering (i.e., interpretability) along the remaining dimensions (Zobnin, 2017). Dufter and Schütze (2019) and Rothe and Schütze (2016) use orthogonal transformations to align a linguistic signal (e.g., a collection of words) to an embedding dimension to obtain an interpretable subspace. However, this method has only been demonstrated in a low-dimensional subspace to date, so its performance in higher dimensional subspaces remains unclear. In a concurrent, independent study (Mathew et al., 2020), the transformation method *POLAR* was proposed to map an existing embedding space to a polar space where each embedding dimension corresponds to a pair of antonyms (i.e., polar opposites). In a recent study (Şenel et al., 2020), an imparting method was proposed in which individual dimensions of the model were aligned with concepts defined a priori based on an external resource. Şenel et al. (2020) demonstrated the effectiveness of this method only for the offline GloVe method, and only the positive direction of each dimension was matched up with a concept.

### 2.2. Gender bias

Ensuring the fairness of mathematical models is one of the most crucial issues in machine learning based information processing. The roles of machine learning and artificial intelligence have an increasing momentum in many real-world applications such as job hiring, granting loans, college applications (Makhlouf et al., 2021). Therefore, algorithms, model parameters, or model features must not include gender, race, ethnic or any other unwanted bias. In Makhlouf et al. (2021), important notions of fairness related to real-world scenarios are extracted, and necessary fairness notions are recommended for each specific setup that includes machine learning.

Bolukbasi et al. (2016) is one of the pioneering studies that investigate gender bias in word embeddings. Authors realize that some occupations that are supposed to be gender-neutral are mapped in favor of one gender by word embeddings. For example, word `man` is closer to `programmer` than `woman` in semantic space. To eliminate this problem, the authors propose two different debiasing methods named soft debiasing and hard debiasing, respectively (Bolukbasi et al., 2016). Caliskan et al. (2017) show that training datasets can unintentionally involve not only gender bias but also morally neutral biases. They also propose the Word Embedding Association Test (WEAT) and the Word Embedding Factual Association Test (WEFAT) to quantify the bias present in texts. In the diachronic study of Garg et al. (2018), it is shown that word embeddings that are trained on different texts from different timelines can reflect social, demographic, and cultural features of the corresponding period. On the other hand, Gonen and Goldberg (2019) approach this issue in a critical way by claiming that the proposed debiasing methods in the literature are not sufficient to remove the bias completely, and that the debiasing methods provide superficial cleaning, and this problem should be dealt with in-depth. The recent advances come with the requirement of detecting and removing biases in contextualized word embeddings and language models. To this end, Liang et al. (2020) propose SENT-DEBIAS method that reduces the social biases in sentence level representations. The proposed method is performed in BERT and ELMO models as an extension of hard debasing in Bolukbasi et al. (2016).

Gender bias is not limited to exist only in word embeddings. Recommender systems and search engines also host gender bias in various ways. Melchiorre et al. (2021) investigate the gender fairness in recommendation algorithms in the music domain. The authors demonstrate the gender inequality in the recommendation performance in favor of the male user group. In addition, they also show that applying debiasing algorithms are beneficial for the improvement of gender fairness. On the other hand, Fabris et al. (2020) propose a measure named 'Gender Stereotype Reinforcement' to evaluate the tendency of search engines to support gender stereotypes. The effect of the embedding debiasing methods on search engines is also inspected.

Detecting gender discrimination is also as important as eliminating gender bias. There exist many kinds of hate speech in social media (Kocoń et al., 2021). Identifying such expressions that contain hatred and biased patterns is also a significant subject of information processing. For instance, Pamungkas et al. (2020) present a review of the state-of-the-art misogyny detection. The most predictive language features for distinguishing hatred and biased content are also presented. Learning these features takes an important part in both detecting and eliminating gender bias in machine learning-based information processing models.

## 3. Research objectives

Our main contributions and research objectives are as follows:

- We propose BiImp, a bidirectional imparting algorithm to improve interpretability of word embeddings that utilizes both directions of each embedding dimension separately to encode different concepts.
- We demonstrate that the bidirectional imparting of arbitrary concepts offers superior performance compared to encoding of polar opposites to each embedding dimension, in terms of interpretability, intrinsic and downstream evaluation tasks.
- We perform comprehensive evaluations and provide comparison with previous work, showing that BiImp achieves greater interpretability without sacrificing performance.
- We propose for the first time an imparting method to concentrate gender information to a designated embedding dimension, along with an hybrid method that achieves concurrent gender and interpretability imparting. We show that this dimension effectively captures gender information and improves the performance of gender debiasing methods, in terms of gender bias metrics and high-level evaluation tasks.

## 4. Methods

### 4.1. Imparting

Unidirectional imparting (*UniImp*) is a method that enhances interpretability in GloVe word embeddings by forcing words related to predefined concepts to project more strongly onto individual embedding dimensions (Şenel et al., 2020). To achieve this, GloVe's cost is modified as follows:

$$\begin{aligned}\sum_{i,j=1}^{V} f(X_{ij}) \Bigg[ & \left(\vec{w}_i^T \vec{\tilde{w}}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2 \\ + \ & k^g \left( \sum_{c=1}^{C} \mathbb{1}_{i\in F_c}\ g(w_{i,c}) + \sum_{c=1}^{C} \mathbb{1}_{j\in F_c}\ g(\tilde{w}_{j,c}) \right) \Bigg]\end{aligned} \tag{1}$$

where $\vec{w}_i$ and $\vec{\tilde{w}}_j$ denote word and context vectors, $w_{i,c}$ and $\tilde{w}_{j,c}$ denote the $c$th components of word and context vectors, $b_i$ and $\tilde{b}_j$ denote word and context biases, $X_{ij}$ denotes co-occurrence of the $i$th and $j$th words in the vocabulary, $V$ denotes vocabulary size, and $f(\cdot)$ is a weighting function to prevent bias from rare words. The first term in the cost is GloVe's original cost function. It aims to capture semantic structure in the embedding model based on word co-occurrences. The second term aims to align embedding dimensions with word-groups. In this latter term, $C$ denotes the number of word-groups ($C \leq dim(\vec{w})$), $\mathbb{1}_{x\in S}$ is the indicator variable for the inclusion $x \in S$, $F_c$ denotes the indices of words that belong to the $c$th group, $k_g$ controls the relative weighting of the second term, and $g(\cdot)$ is a monotone decreasing function that adjusts the size of the updates during training. $g(\cdot)$ is defined as:

$$g(x) = \begin{cases} 1/2 \cdot exp(-2x), & \text{if } x < 0.5 \\ 1/(4ex), & \text{otherwise.} \end{cases}$$

### *4.2. Generalized bidirectional imparting*

In this paper, we propose BiImp, a generalized imparting framework that is capable of online learning and bidirectional imparting. To alleviate computation and memory limitations, we focus on the skip-gram model of word2vec with negative sampling. The objective that the skip-gram model aims to maximize for a word pair $(i, j)$ is given as:

$$\log\ \sigma(\vec{\tilde{w}}_j{}^T \vec{w}_i) + \sum_{t=1}^{m} \mathbb{E}_{z_t\sim P_n(w)} \left[ log\ \sigma(-\vec{\tilde{w}}_{z_t}{}^T \vec{w}_i) \right]. \tag{2}$$

Although the learning mechanisms of GloVe and word2vec are different, unidirectional imparting can still be implemented by maximizing the following modified objective:

$$\begin{aligned}&\log\ \sigma(\vec{\tilde{w}}_j{}^T \vec{w}_i) + \sum_{t=1}^{m} \mathbb{E}_{z_t\sim P_n(w)} \left[ \log\ \sigma(-\vec{\tilde{w}}_{z_t}{}^T \vec{w}_i) \right] \\ &\quad - k^w \Big( \sum_{c=1}^{C} \mathbb{1}_{i\in F_c}\ g(w_{i,c}) + \sum_{c=1}^{C} \mathbb{1}_{j\in F_c}\ g(\tilde{w}_{j,c}) \Big).\end{aligned} \tag{3}$$

In objectives (2) and (3), $\sigma$ is the sigmoid function, $m$ is number of negative samples and $P_n(w)$ is the unigram distribution ($U(w)$) raised to the power 3/4, and $z_t$ is the index of the word from the $t$th draw from the unigram word distribution. Although the additional terms in (1) and (3) look identical, throughout the training process, their relative influence over the original embedding loss can be significantly different. To account for these differences, different weighting factors $k^g$ and $k^w$ are defined.

Imparting was previously only performed for the positive direction of embedding dimensions. But negative directions are equally suitable to encode semantic, interpretable concepts. Based on this argument, we extend the imparting method to both directions of the embedding dimensions. Given a fixed number for embedding dimensions, BiImp doubles the concept capacity compared to the unidirectional case. Moreover, by aligning opposite concepts such as *good* and *bad* or *male* and *female* with opposing directions of the same dimension, these concepts can be represented in a continuum.

The proposed objective for BiImp, the bidirectionally imparted word2vec model is as follows:

$$\begin{aligned}&\log\ \sigma(\vec{\tilde{w}}_j{}^T \vec{w}_i) + \sum_{t=1}^{m} \mathbb{E}_{z_t\sim P_n(w)} \left[ \log\ \sigma(-\vec{\tilde{w}}_{z_t}{}^T \vec{w}_i) \right] \\ &- k^w \ \Big( \sum_{c=1}^{C^+} \mathbb{1}_{i\in F_c^+}\ g(w_{i,c}) + \sum_{c=1}^{C^+} \mathbb{1}_{j\in F_c^+}\ g(\tilde{w}_{j,c}) \\ &\quad - \sum_{c=1}^{C^-} \mathbb{1}_{i\in F_c^-}\ g(w_{i,c}) - \sum_{c=1}^{C^-} \mathbb{1}_{j\in F_c^-}\ g(\tilde{w}_{j,c}) \ \Big)\end{aligned} \tag{4}$$

where $C^+$ and $C^-$ are the number of word-groups associated with positive and negative directions respectively ($C^+ \leq dim(\vec{w})$, $C^- \leq dim(\vec{w})$). $F_c^+$ and $F_c^-$ denote the indices of words that belong to the $c$th group in the positive and negative directions, respectively.

Here word-groups encoded in opposing directions of a given dimension are referred to as word–group pairs. Ideally, the word–group pairs should not contain overlapping words ($F_c^+ \cap F_c^- = \emptyset\ \ \forall c$) to prevent weak word representations. In practice, this problem can be alleviated by rearrangement of word–group pairs. In this study, we apply the following simple rearrangement procedure to prevent overlap. For a given embedding dimension, we first select two random word-groups. When overlap is present, the second word–group is reselected from the set of remaining unpaired word-groups. This procedure is iterated until all word-groups are paired.[1]

### 4.3. Lexical resources

The imparting method requires an external lexical resource that constitutes a basis for interpretability. A trivial interpretation of an embedding model is possible if each embedding dimension is aligned with a distinct concept, (i.e., a word-group). Since practical embedding models can have variable dimensionality, a broad lexical resource that can be used to flexibly extract an arbitrary number of concepts is desirable. To this end, we utilized two lexical resources which are the Roget's Thesaurus (Roget, 2008) and the WordNet (Miller, 1995).

In Şenel et al. (2020), Roget's Thesaurus is utilized as an external resource. Roget's Thesaurus follows a tree structure, where the actual words and phrases are grouped under 1,000 categories making the leaves of the tree structure. We extract word-groups from the thesaurus by partitioning the tree structure starting at the root node from which all other nodes descend. A threshold $\lambda^r_{max}$ is set for the maximum size of a node. Size of a given node is defined as the number of unique descendant words. During partitioning, each node with size less than the threshold is selected to define a word–group, which consists of descendant words for that node. For an above-threshold node without any children nodes, the word–group was defined as the $\lambda^r_{max}$ descendant words with the highest-frequency ranks. Among the resulting word-groups, the ones that contain less than $\lambda^r_{min}$ words are discarded. Finally, word-groups are constructed after discarding the groups with the largest median frequency ranks (i.e., groups that contain more rare words on average).

In addition to the Roget's Thesaurus, we investigate another important lexical resource that can be used to extract semantic word-groups, the WordNet (Miller, 1995). WordNet is a popular lexical database for English in which nouns, verbs, adjectives and adverbs are grouped together into synsets. Each synset expresses a distinct concept. Synsets are interlinked based on their semantic and lexical relations creating a network of related words and concepts. WordNet is similar to a thesaurus since it can be used to group words together based on meaning. However, there are two important differences between WordNet and a thesaurus. First, the network in WordNet is not based on word forms (i.e., sequence of letters) but on specific senses of words. As such, different senses of a word are represented by different synsets providing semantic disambiguation. Second, semantic relations between words are labeled in WordNet to describe the relation types, unlike a thesaurus where words are grouped merely based on similarity in meaning. WordNet is a comprehensive lexical resource containing 117,000 synsets each of which is linked to other synsets. The most frequently encoded relation between synsets is the super-subordinate relation (also known as hyper-hyponymy) that links more general synsets like *furniture* to increasingly specific synsets like *bed* and *bunkbed*. In other words, the category *furniture* includes *bed*, the category *bed* includes *bunkbed* and so on. In the hierarchical structure of WordNet, all noun synsets ultimately go up the root node *entity*.

### 4.4. Interpretability evaluation

Following Şenel et al. (2020), we evaluate the interpretability of the word embeddings based on SEMCAT categories (Şenel, Utlu et al., 2018) and subcategories (Şenel, Yücesoy et al., 2018). SEMCAT (sub)categories are taken as an approximation for the semantic concepts that humans can use to interpret embedding dimensions. Based on SEMCAT, we calculate the *Interpretability Score IS*, which is a measure of how strongly these (sub)categories are represented in embedding dimensions. This metric is low-cost, fast, reproducible and was shown to correlate well with human judgement (Şenel et al., 2020). However, it cannot capture the difference between interpretability changes in the positive and negative directions of an embedding dimension because it performs maximum pooling over the opposite directions of each dimension. To capture this information, we propose a new directional interpretability score:

$$\begin{aligned}
IS^+_{l,k} &= \max_{n_{min} \le n \le n_k} \frac{|S_k \cap V^+_l(\lambda \times n)|}{n} \times 100 \\
IS^-_{l,k} &= \max_{n_{min} \le n \le n_k} \frac{|S_k \cap V^-_l(\lambda \times n)|}{n} \times 100 \\
IS^+_l &= \max_k IS^+_{l,k}, \quad IS^-_l = \max_k IS^-_{l,k}, \\
IS^+ &= \frac{1}{D} \sum_{l=1}^{D} IS^+_l, \quad IS^- = \frac{1}{D} \sum_{l=1}^{D} IS^-_l
\end{aligned} \tag{5}$$

In Eq. (5), $IS^+_{l,k}$ and $IS^-_{l,k}$ represent the interpretability scores in the positive and negative directions of the $l$th dimension ($l \in \{1, 2, \ldots, D\}$, $D = dim(\vec{w})$) for the $k$th category ($k \in \{1, 2, \ldots, K\}$, $K = 110$) in SEMCAT, respectively. $S_k$ is the set of words in the $k$th category in SEMCAT and $n_k$ is the number of words in $S_k$. $n_{min}$ is the minimum number of words required to construct a semantic category (i.e., to represent a concept). $V_i(\lambda \times n)$ represents the set of $\lambda \times n$ words that have the highest ($V^+_l$) and lowest ($V^-_l$) values in the $l$th dimension of the embedding space. For all evaluations we use $\lambda = 5$.

[1] For cases when word-groups have a substantial proportion of overlapping words, more sophisticated matching algorithms might be necessary. However, here, we were able to find a non-overlapping pairing after a few trials (less than 5).

## 4.5. Gender bias

### 4.5.1. Intrinsic bias evaluation

BiImp matches each dimension with concepts and thereby makes it interpretable: it now clearly represents specific concepts. As Dufter and Schütze (2019) argue, this important property can facilitate removal of unwanted information from the model. A common example of such undesirable information is the inherent gender bias in corpora that is reflected in learned embedding models. Bolukbasi et al. (2016) report that embedding models often contain gender bias, particularly for occupation related words.

As discussed in Section 4.2, an important advantage of BiImp over unidirectional imparting is that two concepts with opposite meanings can be represented in a single dimension as a continuum. Since the concepts *male* and *female* are opposites, they can be encoded in the opposite directions of the same dimension, creating a continuous gender dimension. The gender components of words can then be inferred directly from their projections onto the gender dimension. To create a gender dimension, we construct two word-groups corresponding to *male* and *female* concepts using (Bolukbasi et al., 2016)'s gender-specific word set $S$ of 291 professions.

Bolukbasi et al. (2016) proposed two different measures to assess level of gender bias in word embeddings, namely direct bias and indirect bias. Here, we use the direct bias measure:

$$b_{\kappa}^{direct} = \frac{1}{|N|} \sum_{w \in N} |\cos(\vec{w}, \vec{g})|^{\kappa} \tag{6}$$

where $N$ is the set of gender neutral words, $\vec{g} = \vec{w}_{she} - \vec{w}_{he}$ is the gender vector and $\kappa$ is a parameter that controls the relative weighting of high vs. low bias levels; we set $\kappa = 1$. Gender neutral words were obtained by taking the complement of Bolukbasi et al. (2016)'s gender-specific word set $S$ such that $N = W \backslash S$ where $W$ is the set of all words.

To evaluate BiImp on gender bias, we use the stereotypical gender bias levels $b^s$ provided by Bolukbasi et al. (2016) for $S$ (the set of 291 profession words), which were obtained by human assessment.[2] We calculate the correlation $B^g$ between stereotypical biases $b^s$ and the biases $b^{direct}$ based on Eq. (6):

$$B^g = \mathrm{corr}(b^s, b^{direct}) \tag{7}$$

as well as the correlation $B^{gd}$ between the stereotypical biases $b^s$ and the biases $b^{gd}$ from the gender dimension:

$$B^{gd} = \mathrm{corr}(b^s, b^{gd}) \tag{8}$$

where $b^{gd}$ is calculated as:

$$b_p^{gd} = \begin{cases} \min\left(1, \frac{w_p}{\mu_m}\right) & \text{if } w_p \geq 0, \\ -\min\left(1, \frac{w_p}{\mu_f}\right) & \text{if } w_p < 0, \end{cases} \tag{9}$$

$\mu_m$ and $\mu_f$ are the average values of the words in the *male* and *female* word-groups in the gender dimension ($gd$), respectively. $w_p$ stands for the value of the $p$th profession in the gender dimension. The intuition behind Eq. (9) is that we want a value between −1 and 1 (the range of $b^s$) to indicate level of bias. We could map the entire range of values on the dimension to the interval $[-1, 1]$, but that would give too much weight to outliers. We therefore use $\mu_m/\mu_f$ as upper/lower bounds for $w_p$. $B^{gd}$ (resp. $B^g$) indicate how well the BiImp gender dimension (resp. the gender vector $\vec{g}$) captures stereotypical gender bias.

### 4.5.2. Reducing gender bias

Bolukbasi et al. (2016) proposed two methods for gender debiasing: namely *hard debiasing (neutralize and equalize)*, and *soft bias correction*. Here, we consider the hard debiasing method, where the gender subspace is first identified via the principal component analysis (PCA). To do this, difference between word vectors of 10 pairs of gender words (i.e., *female–male, she–he, girl–boy*, etc.) were computed, and PCA was then performed on these 10 difference vectors. The principal component with the largest eigenvalue predominantly captures variance among the difference vectors (around 60% of total variance), suggesting that gender bias primarily lies along a single direction in the embedding space. In the *neutralize* stage, vectors for the gender-neutral words are updated to ensure that their projections onto the first principal component (i.e., gender subspace) is zero. Equality sets are then defined where each set contains a gender pair such as *{men, women}*. In the *equalize* stage, vectors of the words in the equality sets are updated such that the gender pair in each set becomes equidistant to the gender subspace. Therefore, following the equalization stage, each gender-neutral word becomes equidistant to both *men* and *women* vectors.

In this work, we investigate the effect of concentrating gender information in a single dimension of the embedding model via bidirectional imparting. We employ a two-stage approach for reducing gender bias in imparted embedding models. First, we remove the gender dimension from the embedding model to cancel out gender bias as suggested in Dufter and Schütze (2019). Since creation of a gender dimension concentrates gender information in a single dimension, removal of this dimension is expected to remove gender bias from the entire model. Next, we perform hard debiasing as described in Bolukbasi et al. (2016) on the reduced embedding model. Quantitative comparisons of bias level are performed on imparted and reduced embedding models both prior to and after debiasing procedures.

[2] https://github.com/tolga-b/debiaswe/blob/master/data/professions.json Professions that were not in our vocabulary were filtered out.

**Table 1**
Summary statistics of the word–group datasets.

| Word counts | Roget's Thesaurus | | WordNet | |
|---|---|---|---|---|
| | (300 grp.) | (600 grp.) | (300 grp.) | (600 grp.) |
| Total | 20 978 | 40 350 | 26 964 | 18 965 |
| Unique | 12 289 | 19 870 | 18 123 | 13 853 |
| Average | 69.9 ± 53.7 | 67.3 ± 54.6 | 89.9 ± 74.2 | 31.6 ± 15.9 |

### *4.5.3. Bias in classification*

Prost et al. (2019) argue that lower gender bias levels as measured by Eq. (6) do not always translate to reduced gender bias in classification. We therefore also evaluate on *BiosBias* (De-Arteaga et al., 2019), a classification dataset of 397,907 biographies extracted from CommonCrawl. Each biography is annotated as male or female and as being one of 28 different occupations. The task is to classify each subject's occupation given their biography. The train/dev/test split is 258,640/39,790/99,477.

For occupation classification based on an embedding model, single words in a given biography are first projected to the embedding space. Each biography is thereby represented as the average vector of words within the biography. A linear classifier with softmax output is used, and hyperparameters are tuned based on validation set performance. Classification accuracy is used as the performance measure. As a latent measure of gender bias in embedding models, fairness of the classifier to the two genders are examined as described in Hardt et al. (2016) as equality of opportunity. Specifically, we measure the True Positive Rate Gender Gap ($\mathrm{TPR_{gap}}$) and True Negative Rate Gender Gap ($\mathrm{TNR_{gap}}$) for the classifier. $\mathrm{TPR_{gap}}$ for a given occupation is measured as:

$$\begin{aligned}\mathrm{TPR}_{o,\mathrm{gap}} = |Pr\{\hat{B}_o = 1|B_o = 1, B_g = f\}- \\ Pr\{\hat{B}_o = 1|B_o = 1, B_g = m\}|,\end{aligned} \tag{10}$$

where $o$ is an occupation, $B_o$ ($\hat{B}_o$) is the (estimated) occupation of a biography and $B_g$ its gender ($m/f$ = male/female). $\mathrm{TPR_{gap}}$ (resp. $\mathrm{TNR_{gap}}$) is the difference in accuracy between the two genders of detecting the presence (resp. absence) of an occupation. We interpret this as a measure of the gender fairness of the word embeddings for $o$. We compute $\mathrm{TPR_{gap}}/\mathrm{TNR_{gap}}$ as the average over all $\mathrm{TPR}_{o,\mathrm{gap}}/\mathrm{TNR}_{o,\mathrm{gap}}$.

## 5. Experiments and results[3]

In this section, we describe our experiments and present our findings. Section 5.1 describes how we extract word groups from lexical resources. Section 5.2 describes our main experiments for improving interpretability and presents our findings. Section 5.3 presents our gender debiasing experiments. Section 5.4 evaluates the performance of gender de-biased embeddings, and Section 5.5 presents a hybrid gender and interpretability imparted model.

### *5.1. Word–group extraction*

We investigate two lexical resources to extract word groups for imparting: Roget's Thesaurus (Roget, 2008) and WordNet (Miller, 1995). To extract word groups from Roget's Thesaurus, we follow the extraction procedure in Şenel et al. (2020) and extract 300 and 600 word groups by taking $\lambda^w_{min} = 20$ and $\lambda^w_{min} = 15$, respectively. To extract word-groups from WordNet, we follow a similar procedure and partition the hierarchical structure starting from the root node. We follow an iterative approach, where the largest node is divided to its hyponyms in each iteration. Node size is taken as the number of unique words descending from a node after filtering based on the vocabulary extracted from Wikipedia. We discard the nodes with size less than $\lambda^w_{min}$. Iterations are stopped when the number of nodes exceeds the desired word–group count. Note that the desired word-group count may not be achieved if $\lambda^w_{min}$ is selected too large. The groups with the smallest number of member words are discarded to achieve desired word-group. We take $\lambda^w_{min} = 25$ and $\lambda^w_{min} = 15$ for 300 and 600 WordNet word groups, respectively. Table 1 summarizes the statistics for the constructed word-groups.

### *5.2. Interpretability enhancement*

Our training corpus is the English Wikipedia. To pre-process the Wikipedia dump, all document numbers, URLs, HTML syntax and non-alphanumeric characters are cleared. Remaining words are lower-cased. Resulting corpus consists of 2,127,511,369 tokens. Words with less than 100 occurrences are discarded from the corpus. The final vocabulary contains 229,922 unique words (types). To test generalizability of imparting approach, using the 300 Roget word groups and the objectives Eq. (1) and (3), we train two sets of 300-dimensional unidirectionally imparted embeddings (one for GloVe one for word2vec) for different $k^g$ and $k^w$ values. We measure their interpretability using $\mathrm{IS}^+$ (Eq. (5)). Fig. 2 shows interpretability scores for unidirectionally imparted GloVe and word2vec in the positive direction for $n_{min} = 5$ and $n_{min} = 10$. These results suggest that regularization term for imparting is viable for word2vec algorithm as well. However, original word2vec embeddings have lower interpretability values than original GloVe embeddings and word2vec requires stronger regularization than GloVe ($k^w > k^g$) to achieve similar interpretability.

[3] Data and codes are provided at: https://github.com/lksenel/biimp.

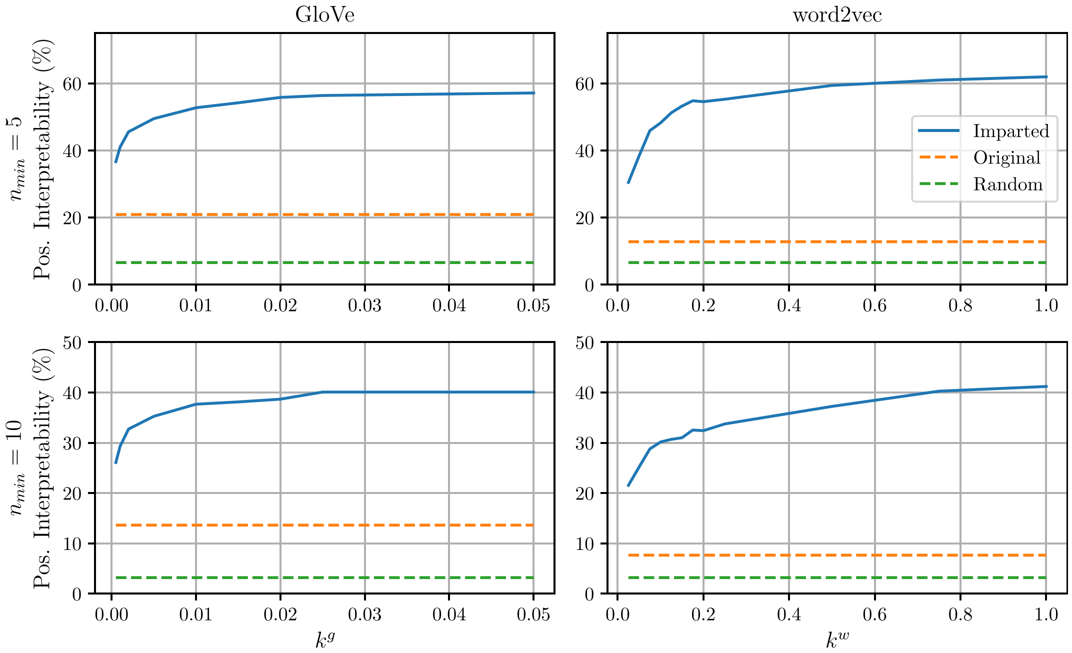

**Fig. 2.** Interpretability scores in the positive direction ($IS^+$) using $n_{min} = 5$ (top row) and $n_{min} = 10$ (bottom row) for unidirectionally imparted GloVe (left column) and word2vec (right column) algorithms for $k^g \in [0.0005, 0.05]$ and $k^w \in [0.025, 1.00]$, respectively. Interpretability scores for original embeddings and a random baseline are displayed for comparison as orange and green dashed lines, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Then, using the 600 word-groups from Roget's Thesaurus and WordNet, we investigate the viability of bidirectional imparting for Word2Vec. Using the objective Eq. (4), we train two sets of 300-dimensional BiImp vectors (one for Roget's and WordNet each) for different $k^w$ values. We additionally train word2vec vectors without bidirectional imparting. For the training all imparted models and the original word2vec model, we use `VOCAB_MIN_COUNT` = 100, `MAX_ITER` = 15, `WINDOW_SIZE` = 8, `NEGATIVE` = 15, `SAMPLE` = $10^{-4}$.

We evaluate the resulting embeddings on two measures: interpretability scores $\mathrm{IS}^+$ and $\mathrm{IS}^-$ (Eq. (5)) and intrinsic performance, based on word similarity[4] (Faruqui & Dyer, 2014) and word analogy[5] (Mikolov, Corrado et al., 2013) tests. Fig. 3 shows interpretability values of the unidirectionally and bidirectionally imparted word2vec embeddings using Roget and WordNet word-groups for $n_{min} = 5$ and $n_{min} = 10$ in both of the positive and negative directions. Bidirectional imparting achieves considerably improved interpretability compared to unidirectional imparting in the negative direction with minimal compromise in the positive direction.

Fig. 4 presents the performances of the embeddings on word similarity and word analogy tests. Performance decreases with increasing $k^w$. However, for bidirectional imparting of WordNet word-groups, performance is on par with original embeddings for $k^w \leq 0.2$. While WordNet word-groups somewhat reduce interpretability compared to Roget word-groups in bidirectional setting, they are much better at preserving the semantic structure of the embedding space as suggested by similarity and analogy tests. Taken together, results in Figs. 3 and 4 suggest that bidirectional imparting of WordNet word-groups at relatively low $k_w$ is the optimal setting for word2vec. Therefore, we use WordNet-based BiImp in the rest of the paper.

### *5.2.1. Interpretability comparison*

We compare BiImp with six state-of-the-art methods for interpretability enhancement: OIWE-IPG (Luo et al., 2015), SOV (Faruqui et al., 2015), Parsimax (Park et al., 2017), Word2Sense (Panigrahi et al., 2019) POLAR (Mathew et al., 2020) and UniImp (Şenel et al., 2020). We do not consider SPINE (Subramanian et al., 2018) because it scaled poorly for large vocabularies in our experiments.

OIWE-IPG was trained on the same corpus as the word2vec embeddings using the default parameters reported in Luo et al. (2015), yielding 300 dimensional vectors. SOV and Parsimax that work on pretrained embeddings were performed on the original word2vec embeddings, again using suggested parameters in Faruqui et al. (2015) and Park et al. (2017), resulting in 1000 and 300 dimensional vectors, respectively. For Word2Sense, we used the publicly available 2250 dimensional pretrained vectors[6] due to computational

[4] Word similarity results were averaged across 13 datasets: WS-353-ALL, SIMLEX-999, VERB-143, SimVerb-3500, WS-353-REL, RW-STANFORD, YP-130, MEN-TR-3k, RG-65, MTurk-771, WS-353-SIM, MC-30, MTurk-287.

[5] http://download.tensorflow.org/data/questions-words.txt.

[6] https://github.com/abhishekpanigrahi1996/Word2Sense.

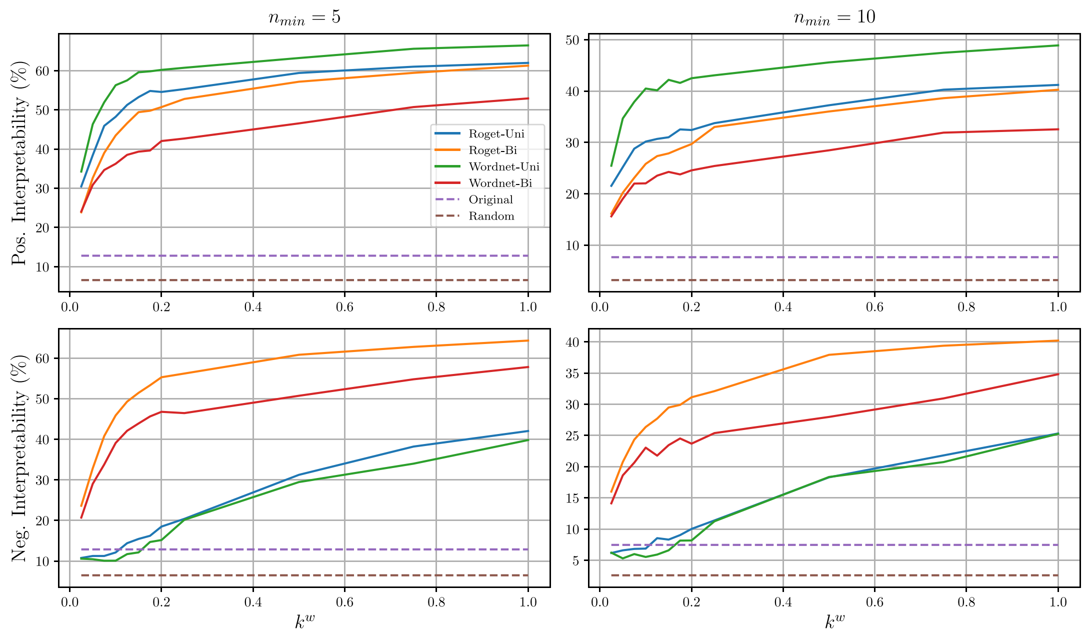

**Fig. 3.** Positive (top) and negative (bottom) direction interpretability scores for unidirectionally imparted word2vec embeddings using Roget's Thesaurus (Roget-Uni) and WordNet (WordNet-Uni) and their bidirectionally imparted versions (BiImp (Roget), BiImp (WordNet)) for $k^w \in [0.025, 1.00]$ along with the original word2vec embedding and a random baseline for $n_{min} = 5$ (left) and $n_{min} = 10$ (right).

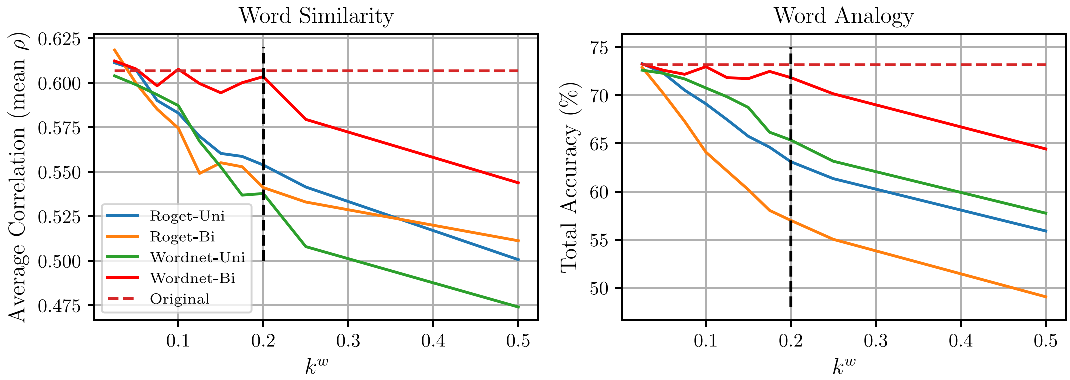

**Fig. 4.** Performance of unidirectionally imparted word2vec embeddings using Roget's Thesaurus (Roget-Uni) and WordNet (WordNet-Uni) and their bidirectionally imparted versions (Roget-Bi, WordNet-Bi) for $k^w \in [0.025, 0.500]$ along with the original word2vec embedding on word similarity (left) and word analogy (right) tests. Word similarity results are presented as the average correlations from 13 different word similarity test sets.

restrictions. For POLAR, we trained two different versions. First, we obtained 1465 dimensional POLAR-large embeddings that were reported in Mathew et al. (2020), by applying polar transformation on Google's pretrained word2vec embeddings[7] using all 1465 antonym pairs. Note that these embeddings were originally trained on a much larger corpus (Google News) with a substantially larger vocabulary (3 million) than our word2vec embeddings. Therefore, POLAR-large embeddings are considerably more expensive than our imparted embeddings in terms of computational and linguistic resources. Second, we obtained 500 dimensional POLAR-small embeddings that are more comparable to imparted embeddings in terms of model dimensionality and resource usage, by performing the polar transformation on our original word2vec embeddings using the default parameters.[8] UniImp embeddings are trained on English Wikipedia (same as BiImp) using Eq. (1) ($k^g = 0.1$ as suggested in Şenel et al. (2020)) and 300 word-groups extracted from Roget's Thesaurus.

Table 2 presents interpretability scores of BiImp for $k^w \in \{0.1, 0.2, 1\}$, OIWE-IPG, SOV, Parsimax, Word2Sense, POLAR$_{small}$, POLAR$_{large}$ and UniImp along with the original word2vec embeddings in positive and negative directions separately for $n_{min} = 5$.

[7] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM.

[8] https://github.com/Sandipan99/POLAR.

**Table 2**
Interpretability scores (cf. Eq. (5), $n_{min} = 5$) of BiImp are higher than all baselines.

| Embedding | Size | Interpretability | |
|---|---|---|---|
| | | pos. | neg. |
| word2vec | 300 | 12.80 | 12.88 |
| OIWE-IPG | 300 | 35.50 | – |
| SOV | 1000 | 14.28 | 13.98 |
| Parsimax | 300 | 18.55 | 17.66 |
| Word2Sense | 2250 | 34.11 | – |
| POLAR$_{small}$ | 500 | 23.89 | 20.8 |
| POLAR$_{large}$ | 1465 | 28.60 | 25.91 |
| UniImp | 300 | 57.49 | 11.38 |
| BiImp$_{k^w=0.1}$ | 300 | 36.24 | 39.10 |
| BiImp$_{k^w=0.2}$ | 300 | 42.04 | 46.77 |
| BiImp$_{k^w=1}$ | 300 | 52.90 | 57.80 |

**Table 3**
Results on the performance evaluation tests. For BiImp, results are averaged across $k^w \in \{0.025, 0.050, \dots, 0.200\}$.

| Task | w2v | IPG | SOV | Parsimax | W2S | POLAR$_s$ | POLAR$_l$ | UniImp | BiImp |
|---|---|---|---|---|---|---|---|---|---|
| Sem. Anlg. | 79.9 | 32.6 | 52.6 | 79.6 | 12.9 | 70.5 | 60.0 | 80.2 | 79.7 |
| Syn. Anlg. | 67.6 | 25.6 | 41.6 | 67.5 | 19.4 | 56.1 | 70.8 | 63.4 | 66.3 |
| Word Sim. | 60.7 | 48.6 | 56.1 | 60.7 | 57.0 | 54.9 | 60.0 | 56.9 | 60.3 |
| Sent. Anly. | 80.3 | 74.5 | 81.8 | 80.3 | 81.2 | 79.1 | 81.8 | 79.0 | 80.00 |
| Quest. Clf. | 85.8 | 79.0 | 87.8 | 85.8 | 77.2 | 84.6 | 82.4 | 81.0 | 84.9 |
| Sports News | 95.9 | 95.5 | 96.9 | 96.0 | 86.6 | 94.7 | 91.8 | 96.0 | 95.7 |
| Relig. News | 87.0 | 85.8 | 88.6 | 86.9 | 85.1 | 84.1 | 84.9 | 84.9 | 87.4 |
| Comp. News | 81.6 | 78.5 | 86.3 | 81.7 | 73.4 | 77.6 | 72.9 | 80.3 | 80.3 |

Note that non-negative embeddings inherently do not have any interpretability in the negative direction. BiImp embeddings are clearly the most interpretable in the negative direction, even for small $k^w$ ($k^w = 0.1$). For the positive direction, interpretability of BiImp is comparable with OIWE-IPG and Word2Sense and is higher than all baselines except UniImp for small $k^w$. For larger $k^w$, interpretability of BiImp is only slightly lower than that of UniImp.

### *5.2.2. Preservation of semantic structure*

In addition to the intrinsic evaluation, we also evaluate the embeddings on three classification tasks:

- **Sentiment Analysis:** A sentence-level binary classification task using the Stanford Sentiment Treebank consisting of thousands of movie reviews (Socher et al., 2013) and their sentiment scores. The development and training sets in the original dataset were aggregated, and reviews with neutral scores were removed (i.e., scores between 0.4 and 0.6). The resulting dataset contained 7407 training and 1751 test samples.
- **Question Classification (TREC):** A question-level multinomial classification task using the TREC dataset (Li & Roth, 2006) consisting of six different types of questions (person, location, entity, number, description, abbreviation). This dataset consisted of 5452 training and 500 test questions.
- **News Classification:** Following Faruqui et al. (2015), three news-level binary classification tasks were considered using the 20 Newsgroup dataset.[9] The following news topics were considered (training/test sample counts): (1) Religion: atheism vs. christian (1079/716); (2) Sports: baseball vs. hockey (1192/796); (3) Computers: IBM vs. Mac (1162/775).

For these high-level NLP tasks, we took the average of the word vectors in input text (can be a sentence, question or news) as input features and trained an SVM classifier that was tuned using 5 fold cross-validation on the training sets.

Table 3 shows results. For BiImp, results are averaged across $k^w \in \{0.025, 0.050, \dots, 0.200\}$. For analogy and similarity tasks, BiImp, UniImp, Parsimax and word2vec have similar scores, suggesting that BiImp does not reduce the quality of word embeddings while improving interpretability. Both POLAR models perform slightly worse than the original embeddings (except for syntactic analogy and sentiment analysis for POLAR-large). OIWE-IPG, SOV and Word2Sense suffer from considerable performance loss in most cases, implying a reduction in the semantic information captured.

For text classification (last five lines), differences between methods are minor, except for Word2Sense embeddings, which perform poorly on question and news classification. SOV (Faruqui et al., 2015) has the best performance on classification, but recall that it has low interpretability (Table 2). BiImp performs comparably to UniImp, Parsimax and word2vec in all tasks. These results demonstrate that BiImp meets both requirements: interpretability and good task performance.

[9] http://qwone.com/~jason/20Newsgroups.

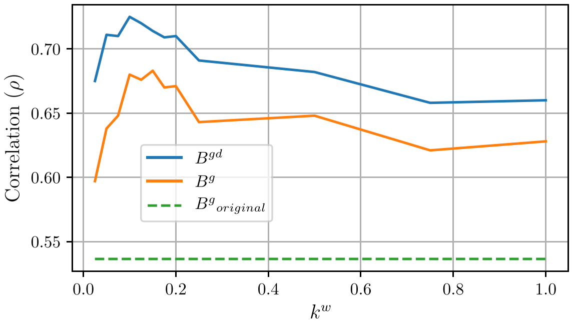

**Fig. 5.** Correlation of human judgments with the gender dimension in BiImp (blue, $B^{gd}$, Eq. (8)), with the gender vector in BiImp (orange, $B^g$, Eq. (7)), and with the gender vector in the original embedding space (dashed green line, $B^g_{original}$). The BiImp gender dimension clearly has the highest correlation with human judgments. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

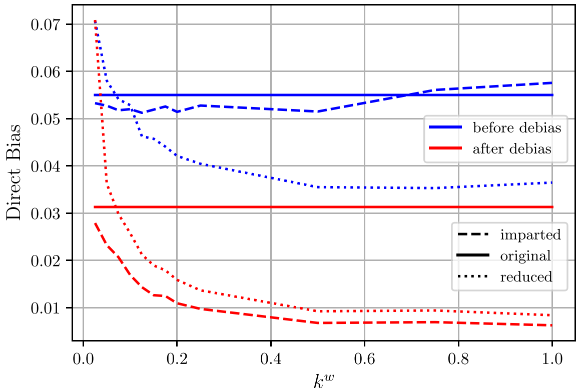

**Fig. 6.** Direct bias ($b_1^{direct}$, see Eq. (6)) of the BiImp (dashed lines) and reduced (dotted lines) embeddings as a function of $k^w$. Solid lines: $b_1^{direct}$ of the original embeddings. Blue/red: Results before/after hard debiasing. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

### 5.3. Gender debiasing

#### 5.3.1. Intrinsic bias

We calculate $B^g$ (Eq. (7)) and $B^{gd}$ (Eq. (8)) for BiImp. Additionally, we calculate $B^g$ for the original word2vec embeddings ($B^g_{original}$). Fig. 5 shows that Pearson's correlation coefficients of human judgments with the BiImp gender dimension (blue, $B^{gd}$, Eq. (8)) is higher than their correlation with the gender vector $\vec{w}_{she} - \vec{w}_{he}$ of BiImp (orange, $B^g$, Eq. (7)) and also higher than the correlation with the original word2vec gender vector (dashed green line). This result suggests that BiImp's gender dimension densely captures gender information. Interestingly, $B^g$ is much higher for BiImp than for the original embeddings ($B^g_{original}$), indicating that BiImp improves the quality of the gender vector as well.

We investigate the effect of (i) gender imparting, (ii) removing the gender dimension from the embeddings (iii) hard debiasing (Bolukbasi et al., 2016) on the gender bias level of the embedding spaces. Specifically, we measure the bias level of the original, imparted and reduced embeddings before and after hard debiasing using Eq. (6). Fig. 6 shows the bias levels. Naturally, imparting a single dimension with gender information does not alter the overall bias in the word embeddings, but rather concentrates most of the bias on a single dimension as implied by Fig. 5. Removing this dimension from the embedding space then considerably reduces the bias, especially for larger $k^w$. After hard debiasing, $b_1^{direct}$ of the full and reduced imparted models (red dashed and dotted lines) are closer, and substantially lower than that of word2vec. These results show that learning an embedding space with an explicit gender dimension enhances the performance of hard debiasing.

#### 5.3.2. Bias in classification

Prost et al. (2019) give evidence that hard debiasing introduces elevated gender bias in high-level classification tasks when compared with the original embedding model. We therefore also use *strong debiasing* (Prost et al., 2019), a method that alleviates this issue by taking $N$ (Eq. (6)) as the entire vocabulary as opposed to just gender neutral words.

Table 4 compares original embeddings, hard debiasing, strong debiasing and the combination of BiImp and strong debiasing (B+S) on accuracy (to measure task performance) and TPR/TNR (Eq. (10), to measure classification fairness). The dataset is BiosBias. Hard debiasing has relatively high TPR/TNR, suggesting it reduces classification fairness. Strong debiasing on original word2vec

**Table 4**
Accuracy and True Positive/Negative Rate (TPR/TNR) on the occupation classification task. B + S = BiImp + strong debiasing.

| Embedding | Acc. | TPR$_{gap}$ | TNR$_{gap}$ |
|---|---|---|---|
| word2vec | .717 | .094 | .0034 |
| Hard debiasing | .700 | .105 | .0037 |
| Strong debiasing | .699 | .087 | .0033 |
| B + S$_{k^w=.1}$ | .697 | .066 | .0022 |
| B + S$_{k^w=.5}$ | .699 | .067 | .0024 |

**Table 5**
Results of embeddings from gender bias experiments on the performance evaluation tests.

| Task | Before debias | | | After debias | | |
|---|---|---|---|---|---|---|
| | word2vec | Imparted | Reduced | word2vec | Imparted | Reduced |
| Sem. Anlg. | 79.87 | 79.00 ± 0.50 | 79.16 ± 0.50 | 78.65 | 78.92 ± 0.57 | 78.99 ± 0.61 |
| Syn. Anlg. | 67.63 | 66.39 ± 0.99 | 66.48 ± 1.01 | 67.46 | 66.42 ± 0.96 | 66.43 ± 1.00 |
| Word Sim. | 60.68 | 60.08 ± 0.66 | 60.21 ± 0.52 | 60.64 | 60.12 ± 0.67 | 60.28 ± 0.53 |
| Sent. Anly. | 80.30 | 79.95 ± 0.36 | 79.94 ± 0.33 | 79.84 | 79.99 ± 0.37 | 79.98 ± 0.41 |
| Quest. Clf. | 85.80 | 84.63 ± 0.59 | 86.00 ± 0.92 | 86.20 | 86.27 ± 0.74 | 86.03 ± 0.80 |
| Sports News | 95.85 | 95.33 ± 0.27 | 95.34 ± 0.25 | 95.10 | 95.33 ± 0.27 | 95.33 ± 0.29 |
| Relig. News | 87.01 | 86.19 ± 0.61 | 86.10 ± 0.57 | 86.03 | 86.24 ± 0.59 | 86.18 ± 0.59 |
| Comp. News | 81.55 | 78.74 ± 0.84 | 78.73 ± 0.99 | 78.84 | 78.68 ± 0.83 | 78.63 ± 0.81 |

**Table 6**
Results of evaluation tests for the hybrid gender and interpretability imparted embeddings.

| Task | $k^w = 0.1$ | $k^w = 0.2$ | $k^w = 1$ |
|---|---|---|---|
| Semantic Anlg. | 79.07 | 78.13 | 73.25 |
| Syntactic Anlg. | 66.61 | 65.17 | 45.58 |
| Word Sim. | 60.62 | 59.11 | 48.94 |
| Sentiment Anly. | 80.41 | 79.55 | 79.84 |
| Question Clf. | 84.60 | 85.00 | 84.20 |
| Sports News | 96.11 | 95.73 | 95.73 |
| Religion News | 85.89 | 87.43 | 88.55 |
| Comput. News | 81.42 | 81.03 | 79.74 |
| Interp.$^{+}_{n_{min}=5}$ | 36.88 | 41.22 | 54.28 |
| Interp.$^{-}_{n_{min}=5}$ | 38.47 | 44.79 | 58.50 |
| Interp$^{+}_{n_{min}=10}$ | 22.41 | 24.17 | 34.07 |
| Interp.$^{-}_{n_{min}=10}$ | 22.49 | 23.89 | 35.43 |
| Gender B.$_{reduced}$ | 0.0470 | 0.0403 | 0.0441 |
| Gender B.$_{debiased}$ | 0.0168 | 0.0122 | 0.0148 |

results in a relatively limited change in classification fairness. Yet when BiImp and strong debiasing are combined (B+S), TPR$_{gap}$ and TNR$_{gap}$ are substantially lowered without a major compromise in accuracy. These results provide further evidence that concentration of gender information on an embedding dimension improves performance of debiasing methods.

### 5.4. Performance of gender biased embeddings

A potential risk of debiasing on gender-imparted models is undesirable loss of semantic structure in the embedding space that might compromise task performance. To rule out this risk, we evaluate the embeddings in the gender-bias experiments on intrinsic tests and downstream classification tasks. For the imparted and reduced embeddings, we averaged the results across $k^w$. Table 5 shows that all the evaluated embeddings perform nearly as good as the original embeddings on all tasks, except a slightly reduced performance on computer news classification task. These results indicate that debiasing of gender-imparted embeddings successfully preserves semantic structure of the embedding space.

### 5.5. Hybrid gender and interpretability imparted embeddings

We demonstrate the feasibility of BiImp for concurrent gender and interpretability imparting. To do this, we obtain a hybrid model where the first dimension was encoded with gender word-groups and the remaining 299 dimensions were bidirectionally imparted with word-groups extracted from WordNet. Evaluation on gender bias, interpretability and task performance were repeated on this hybrid model. Table 6 shows the evaluation results. Hybrid model performs similarly to only WordNet imparted BiImp

(Sections 4.4 and 5.2.2) in interpretability and task performance evaluations, and performs similarly to only gender imparted BiImp (Section 5.3.1) in gender bias evaluations. These results indicate that BiImp enables gender debiasing and interpretability enhancement simultaneously in embedding models without compromising task performance.

## 6. Discussion of results and implications

The implications of the presented results can be organized under three main folds as follows.

- BiImp generates interpretable word embeddings by disclosing the hidden encoded structure of word embedding models without performance degradations on semantic tasks: Producing interpretable word embeddings has a critical role in deciphering the black-box behavior of language models extensively used in NLP-based information processing. Studies generating interpretable embeddings mostly give up some of the semantic properties captured by word vectors. Our experimental results show that BiImp brings interpretable word embeddings without making compromises on the semantic task performances.
- BiImp has a flexibility to be adapted to distinctive learning scenarios and semantic tasks: Aside from the main objective, BiImp is also compatible for different training schemes for word embeddings. BiImp can be easily adapted to both online learning-based and co-occurrence matrix-based training procedures. In addition, different lexical sources can be utilized without any additional cost. One can infer that BiImp presents a large spectrum of interpretable embeddings with a performance at the state-of-the-art level in various tasks ranging from word analogy to text classification.
- BiImp can also be deployed to capture and mitigate any kind of human biases that exist in word embeddings: On the other hand, imparting interpretability to word embeddings enables us to enhance word embeddings in various ways. As shown in the experimental results, capturing human biases in a dimension and removing that dimension lead to better debiasing results. This feature of BiImp embeddings can be extended to other bias types without any difficulty. Furthermore, task or domain-specific interpretable word embeddings can be obtained by adjusting the corresponding word groups assigned to embedding dimensions according to the task or domain. As a result, BiImp offers wide liberty in studying word embeddings without any further computational efforts.

## 7. Conclusion

We introduced BiImp, a new method for enhancing interpretability of word embeddings by bidirectional imparting of concepts extracted from lexical resources. BiImp was implemented for the scalable word2vec algorithm, and semantic concepts were extracted from Roget's Thesaurus and WordNet. In contrast to prior work, BiImp uses both directions along each dimension of the vector space separately, enabling encoding of two different concepts; the two concepts can be chosen arbitrarily or chosen as opposite concepts as a special case. As a result, BiImp makes more efficient use of the embedding space while increasing encoding flexibility.

We showed that BiImp achieves higher interpretability of word embeddings compared to state-of-the-art methods, particularly in the negative direction. At the same time, evaluation on word similarity/analogy tests as well as sentiment, news and question classification showed that BiImp does not sacrifice task performance. Thus, BiImp offers a favorable trade-off between the goals of enhancing interpretability and maintaining task performance.

BiImp represents opposite concepts in a single dimension on a continuum. As an important demonstration, we used BiImp to concentrate gender information in a single gender dimension. We showed that this gender dimension has a high correlation with stereotypical gender bias as measured by human judgments. Furthermore, we showed that this gender dimension is useful for reducing gender bias when coupled with debiasing. The combination of BiImp and debiasing achieved lower levels of gender bias and improved classification fairness. These results highlight the potential of BiImp in reducing biases and stereotypes present in word embeddings.

Here, the imparting method was demonstrated to improve interpretability and reduce gender bias in word2vec embedding models, using concepts from two common lexical sources. That said, imparting through modification of the learning objective is easily adaptable to different embedding algorithms, and to different lexical resources. The imparting framework can also be adopted for goals beyond interpretability enhancement, such as improvement of task performance. If imparting is used to encode task-relevant concepts, similar task performance can be achieved using simpler models with fewer dimensions. In turn, this can offer benefits in terms of memory requirements and computational load.

Lastly, we studied BiImp in the scope of static word embeddings. Extending BiImp to the contextualized word embeddings can be further investigated as a future work.

## CRediT authorship contribution statement

**Lütfi Kerem Şenel:** Data curation, Software, Validation, Formal analysis, Investigation, Visualization, Resources, Writing – original draft, Writing – review & editing. **Furkan Şahinuç:** Data curation, Software, Validation, Formal analysis, Investigation, Visualization, Resources, Writing – original draft. **Veysel Yücesoy:** Validation, Writing – review & editing. **Hinrich Schütze:** Conceptualization, Methodology, Formal analysis, Supervision, Resources, Writing – review & editing, Funding acquisition. **Tolga Çukur:** Conceptualization, Methodology, Formal analysis, Supervision, Resources, Writing – original draft, Writing – review & editing, Funding acquisition. **Aykut Koç:** Conceptualization, Methodology, Formal analysis, Supervision, Resources, Writing – original draft, Writing – review & editing, Funding acquisition.

## Acknowledgments

A. Koç acknowledges the support from TUBITAK 1001 Grant (120E346). T. Cukur acknowledges support from TUBA GEBIP 2015 and BAGEP 2017 fellowships.

## References

Agarwal, O., Durupınar, F., Badler, N. I., & Nenkova, A. (2019). Word embeddings (also) encode human personality stereotypes. In *Proceedings of the Joint Conference on Lexical and Computational Semantics (SEM 2019)* (pp. 205–211). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/S19-1023.

Arora, S., Li, Y., Liang, Y., Ma, T., & Risteski, A. (2018). Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association of Computational Linguistics*, *6*, 483–495. http://dx.doi.org/10.1162/tacl_a_00034.

Bagheri, E., Ensan, F., & Al-Obeidat, F. N. (2018). Neural word and entity embeddings for ad hoc retrieval. *Information Processing & Management*, *54*, 657–673. http://dx.doi.org/10.1016/j.ipm.2018.04.007.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. http://dx.doi.org/10.1162/tacl_a_00051.

Bollegala, D., Mohammed, A., Maehara, T., & Kawarabayashi, K.-i. (2016). Joint Word Representation Learning Using a Corpus and a Semantic Lexicon. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)* (pp. 2690–2696).

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)* (pp. 4356–4364). Curran Associates, Inc..

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. http://dx.doi.org/10.1126/science.aal4230.

Celikyilmaz, A., Hakkani-Tur, D., Pasupat, P., & Sarikaya, R. (2015). Enriching word embeddings using knowledge graph for semantic tagging in conversational dialog systems. In *Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium Series*.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)* (pp. 2180–2188). Curran Associates, Inc..

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., & Kalai, A. T. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *FAT* '19*, *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 120–128). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3287560.3287572.

Demszky, D., Garg, N., Voigt, R., Zou, J., Gentzkow, M., Shapiro, J., & Jurafsky, D. (2019). Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 2970–3005). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1304.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1423.

Dufter, P., & Schütze, H. (2019). Analytical methods for interpretable ultradense word embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 1185–1191). Hong Kong, China: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1111.

Elnagar, A., Al-Debsi, R., & Einea, O. (2020). Arabic text classification using deep learning models. *Information Processing & Management*, *57*(1), Article 102121. http://dx.doi.org/10.1016/j.ipm.2019.102121.

Fabris, A., Purpura, A., Silvello, G., & Susto, G. A. (2020). Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management*, *57*(6), Article 102377. http://dx.doi.org/10.1016/j.ipm.2020.102377.

Faruqui, M., & Dyer, C. (2014). Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 19–24). Baltimore, Maryland: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/P14-5004.

Faruqui, M., Tsvetkov, Y., Yogatama, D., Dyer, C., & Smith, N. A. (2015). Sparse overcomplete word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1491–1500). Beijing, China: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/P15-1144.

Fyshe, A., Talukdar, P. P., Murphy, B., & Mitchell, T. M. (2014). Interpretable semantic vectors from a joint model of brain- and text- based meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 489–499). Baltimore, Maryland: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/P14-1046.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. http://dx.doi.org/10.1073/pnas.1720347115.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 609–614). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1061.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1489–1501). Berlin, Germany: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P16-1141.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)* (pp. 3315–3323). Curran Associates, Inc..

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. http://dx.doi.org/10.1038/nature17637.

Iter, D., Yoon, J., & Jurafsky, D. (2018). Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (pp. 136–146). New Orleans, LA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W18-0615.

Ji, D., Gao, J., Fei, H., Teng, C., & Ren, Y. (2020). A deep neural network model for speakers coreference resolution in legal texts. *Information Processing & Management*, *57*(6), Article 102365. http://dx.doi.org/10.1016/j.ipm.2020.102365.

Ji, D., Tao, P., Fei, H., & Ren, Y. (2020). An end-to-end joint model for evidence information extraction from court record document. *Information Processing & Management*, *57*(6), Article 102305. http://dx.doi.org/10.1016/j.ipm.2020.102305.

Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T., & Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, *58*(5), Article 102643. http://dx.doi.org/10.1016/j.ipm.2021.102643.

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1384–1397). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 302–308). Baltimore, Maryland: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/P14-2050.

Li, X., & Roth, D. (2006). Learning question classifiers: The role of semantic information. *Natural Language Engineering*, *12*(3), 229–249. http://dx.doi.org/10.1017/S1351324905003955.

Liang, P. P., Li, I. M., Zheng, E., Lim, Y. C., Salakhutdinov, R., & Morency, L.-P. (2020). Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5502–5515). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.488.

Liu, Q., Jiang, H., Wei, S., Ling, Z.-H., & Hu, Y. (2015). Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1501–1511). Beijing, China: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/P15-1145.

López-Santillan, R., Montes-Y-Gomez, M., Gonzalez-Gurrola, L. C., Ramirez-Alonso, G., & Prieto-Ordaz, O. (2020). Richer document embeddings for author profiling tasks based on a heuristic search. *Information Processing & Management*, *57*(4), Article 102227. http://dx.doi.org/10.1016/j.ipm.2020.102227.

Luo, H., Liu, Z., Luan, H., & Sun, M. (2015). Online learning of interpretable word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1687–1692). Lisbon, Portugal: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D15-1196.

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021). Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, *58*(5), Article 102642. http://dx.doi.org/10.1016/j.ipm.2021.102642.

Mathew, B., Sikdar, S., Lemmerich, F., & Strohmaier, M. (2020). The POLAR framework: Polar opposites enable interpretability of pre-trained word embeddings. In *Proceedings of the Web Conference 2020* (pp. 1548–1558). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3366423.3380227.

Melchiorre, A. B., Rekabsaz, N., Parada-Cabaleiro, E., Brandl, S., Lesota, O., & Schedl, M. (2021). Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management*, *58*(5), Article 102666. http://dx.doi.org/10.1016/j.ipm.2021.102666.

Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR)* (pp. 1–12).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)* (pp. 3111–3119). Curran Associates, Inc..

Miller, G. A. (1995). WordNet: A lexical database for english. *Communications of the ACM*, *38*(11), 39–41. http://dx.doi.org/10.1145/219717.219748.

Moudjari, L., Benamara, F., & Akli-Astouati, K. (2021). Multi-level embeddings for processing arabic social media contents. *Computer Speech and Language*, *70*, Article 101240. http://dx.doi.org/10.1016/j.csl.2021.101240.

Mrkšić, N., Ó Séaghdha, D., Thomson, B., Gašić, M., Rojas-Barahona, L. M., Su, P.-H., Vandyke, D., Wen, T.-H., & Young, S. (2016). Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 142–148). San Diego, California: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N16-1018.

Mumcuoğlu, E., Öztürk, C. E., Ozaktas, H. M., & Koç, A. (2021). Natural language processing in law: Prediction of outcomes in the higher courts of Turkey. *Information Processing & Management*, *58*(5), Article 102684. http://dx.doi.org/10.1016/j.ipm.2021.102684.

Murphy, B., Talukdar, P., & Mitchell, T. (2012). Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In *Proceedings of International Conference on Computational Linguistics (COLING)* (pp. 1933–1950).

Nozza, D., Manchanda, P., Fersini, E., Palmonari, M., & Messina, E. (2021). Learningtoadapt with word embeddings: Domain adaptation of named entity recognition systems. *Information Processing & Management*, *58*(3), Article 102537. http://dx.doi.org/10.1016/j.ipm.2021.102537.

Pamungkas, E. W., Basile, V., & Patti, V. (2020). Misogyny detection in Twitter: A multilingual and cross-domain study. *Information Processing & Management*, *57*(6), Article 102360. http://dx.doi.org/10.1016/j.ipm.2020.102360.

Pamungkas, E. W., Basile, V., & Patti, V. (2021). A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, *58*(4), Article 102544. http://dx.doi.org/10.1016/j.ipm.2021.102544.

Panigrahi, A., Simhadri, H. V., & Bhattacharyya, C. (2019). Word2Sense: Sparse interpretable word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5692–5705). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1570.

Papagiannopoulou, E., & Tsoumakas, G. (2018). Local word vectors guiding keyphrase extraction. *Information Processing & Management*, *54*(6), 888–902. http://dx.doi.org/10.1016/j.ipm.2018.06.004.

Park, S., Bak, J., & Oh, A. (2017). Rotated word vector representations and their interpretability. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 401–411). Copenhagen, Denmark: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D17-1041.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1162.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, *9*(1), 963. http://dx.doi.org/10.1038/s41467-018-03068-4.

Pronoza, E., Panicheva, P., Koltsova, O., & Rosso, P. (2021). Detecting ethnicity-targeted hate speech in Russian social media texts. *Information Processing & Management*, *58*(6), Article 102674. http://dx.doi.org/10.1016/j.ipm.2021.102674.

Prost, F., Thain, N., & Bolukbasi, T. (2019). Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (pp. 69–75). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-3810.

Qian, Y., Deng, X., Ye, Q., Ma, B., & Yuan, H. (2019). On detecting business event from the headlines and leads of massive online news articles. *Information Processing & Management*, *56*(6), Article 102086. http://dx.doi.org/10.1016/j.ipm.2019.102086.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners: Technical Report*.

Roget, P. M. (2008). *Roget's International Thesaurus, 3/E*. Oxford and IBH Publishing.

Rothe, S., & Schütze, H. (2016). Word embedding calculus in meaningful ultradense subspaces. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computer Linguistics, http://dx.doi.org/10.18653/v1/p16-2083.

Roy, D., Ganguly, D., Mitra, M., & Jones, G. J. (2019). Estimating Gaussian mixture models in the local neighbourhood of embedded word vectors for query performance prediction. *Information Processing & Management*, *56*(3), 1026–1045. http://dx.doi.org/10.1016/j.ipm.2018.10.009.

Roy, P. K., Kumar, A., Singh, J. P., Dwivedi, Y. K., Rana, N. P., & Raman, R. (2021). Disaster related social media content processing for sustainable cities. *Sustainable Cities and Society*, *75*, Article 103363. http://dx.doi.org/10.1016/j.scs.2021.103363.

Ruan, Y.-P., Ling, Z.-H., & Hu, Y. (2016). Exploring semantic representation in brain activity using word embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 669–679). Austin, Texas: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D16-1064.

Şahinuç, F., & Koç, A. (2021). Zipfian regularities in non-point word representations. *Information Processing & Management*, *58*(3), Article 102493. http://dx.doi.org/10.1016/j.ipm.2021.102493.

Schick, T., & Schütze, H. (2020). BERTRAM: Improved word embeddings have big impact on contextualized model performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3996–4007). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.368, URL: https://aclanthology.org/2020.acl-main.368.

Şenel, L. K., Utlu, I., Şahinuç, F., Ozaktas, H. M., & Koç, A. (2020). Imparting interpretability to word embeddings while preserving semantic structure. *Natural Language Engineering*, 1–26. http://dx.doi.org/10.1017/S1351324920000315.

Şenel, L. K., Utlu, I., Yücesoy, V., Koç, A., & Cukur, T. (2018). Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1769–1779. http://dx.doi.org/10.1109/TASLP.2018.2837384.

Şenel, L. K., Yücesoy, V., Koç, A., & Cukur, T. (2018). Interpretability analysis for turkish word embeddings. In *2018 26th Signal Processing and Communications Applications Conference (SIU)* (pp. 1–4). IEEE, http://dx.doi.org/10.1109/SIU.2018.8404244.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1631–1642). Seattle, Washington, USA: Association for Computational Linguistics.

Subramanian, A., Pruthi, D., Jhamtani, H., Berg-Kirkpatrick, T., & Hovy, E. (2018). SPINE: SParse Interpretable Neural Embeddings. In: Proceedings of the Thirty Second AAAI Conference on Artificial Intelligence (AAAI).

Tuke, J., Nguyen, A., Nasim, M., Mellor, D., Wickramasinghe, A., Bean, N., & Mitchell, L. (2020). Pachinko prediction: A Bayesian method for event prediction from social media data. *Information Processing & Management*, *57*(2), Article 102147. http://dx.doi.org/10.1016/j.ipm.2019.102147.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)* (pp. 5998–6008). Curran Associates, Inc..

Voppel, A., de Boer, J., Brederoo, S., Schnack, H., & Sommer, I. (2021). Quantified language connectedness in schizophrenia-spectrum disorders. *Psychiatry Research*, *304*, Article 114130. http://dx.doi.org/10.1016/j.psychres.2021.114130, URL: https://www.sciencedirect.com/science/article/pii/S0165178121004261.

Yang, X., & Mao, K. (2016). Task independent fine tuning for word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(4), 885–894. http://dx.doi.org/10.1109/TASLP.2016.2644863.

Yu, M., & Dredze, M. (2014). Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 545–550). Baltimore, Maryland: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/P14-2089.

Yu, L.-C., Wang, J., Lai, K. R., & Zhang, X. (2017). Refining word embeddings using intensity scores for sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(3), 671–681. http://dx.doi.org/10.1109/TASLP.2017.2788182.

Yüksel, A., Uğurlu, B., & Koç, A. (2021). Semantic change detection with gaussian word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3349–3361. http://dx.doi.org/10.1109/TASLP.2021.3120645.

Zhang, Y., Han, K., Worth, R., & Liu, Z. (2020). Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature communications*, *11*(1), 1–13. http://dx.doi.org/10.1038/s41467-020-15804-w.

Zhao, X., Wang, D., Zhao, Z., Liu, W., Lu, C., & Zhuang, F. (2021). A neural topic model with word vectors and entity vectors for short texts. *Information Processing & Management*, *58*(2), Article 102455. http://dx.doi.org/10.1016/j.ipm.2020.102455.

Zobnin, A. (2017). Rotations and interpretability of word embeddings: The case of the Russian language. In *Proceedings of the International Conference on Analysis of Images, Social Networks and Texts (AIST)* (pp. 116–128). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-73013-4_11.