SexWEs: Domain-Aware Word Embeddings via Cross-Lingual Semantic Specialisation for Chinese Sexism Detection in Social Media

Aiqi Jiang, Arkaitz Zubiaga

Queen Mary University of London
{a.jiang, a.zubiaga}@qmul.ac.uk

Abstract

The goal of sexism detection is to mitigate negative online content targeting certain gender groups of people. However, the limited availability of labeled sexism-related datasets makes it problematic to identify online sexism for low-resource languages. In this paper, we address the task of automatic sexism detection in social media for one low-resource language – Chinese. Rather than collecting new sexism data or building cross-lingual transfer learning models, we develop a cross-lingual domain-aware semantic specialisation system in order to make the most of existing data. Semantic specialisation is a technique for retrofitting pre-trained distributional word vectors by integrating external linguistic knowledge (such as lexico-semantic relations) into the specialised feature space. To do this, we leverage semantic resources for sexism from a high-resource language (English) to specialise pre-trained word vectors in the target language (Chinese) to inject domain knowledge. We demonstrate the benefit of our sexist word embeddings (SexWEs) specialised by our framework via intrinsic evaluation of word similarity and extrinsic evaluation of sexism detection. Compared with other specialisation approaches and Chinese baseline word vectors, our SexWEs shows an average score improvement of 0.033 and 0.064 in both intrinsic and extrinsic evaluations, respectively. The ablative results and visualisation of SexWEs also prove the effectiveness of our framework on retrofitting word vectors in low-resource languages.

Introduction

Due to the volume of incidents, hostile behaviours and violence in social media, manual inspection and moderation have become unmanageable, especially for minorities and minorised communities (Jha and Mamidi 2017; Fersini, Nozza, and Rosso 2020; Rodríguez-Sánchez, Carrillo-de Albornoz, and Plaza 2020). Sex is commonly a sensitive topic, and sexist content is of high subjectivity. The high cognition and tolerance thresholds of hostile gender-biased behaviour by certain gender groups can exacerbate gender-based hatred and violence online (Shi and Zheng 2020). Sexist speech refers to those promoting gender-based abuse and violence against an individual or a gender group of people on actual or perceived aspects of personal characteristics (e.g. physical gender differences) (Jiang et al. 2022), manifested in various behaviours (e.g. stereotyping, ideological issues, and sexual violence) (Manne 2017; Anzovino, Fersini, and Rosso 2018). Glick and Fiske (2001) define sexism as an ambivalent attitude manifested through both hostility and benevolence. Hostile sexism is characterised by an explicitly negative attitude towards gender groups (e.g. misogyny), while benevolent sexism is more subtle with seemingly positive characteristics. Most studies focus more on detecting hostile sexism, overlooking implicit expressions of sexism (Waseem and Hovy 2016; Pamungkas, Basile, and Patti 2020). Hence, mitigating online sexism in a wide spectrum of sexist behaviours is crucial as these are, in fact, extremely dangerous and harmful to society (Richardson-Self 2018).

Research in sexism detection has recently increased in popularity (Rodríguez-Sánchez, Carrillo-de Albornoz, and Plaza 2020; Jiang and Zubiaga 2021). However, sexism-related resources are predominantly available in English (Waseem and Hovy 2016; Jha and Mamidi 2017; Samory et al. 2021) and Indo-European languages (Fersini, Nozza, and Rosso 2020; Chiril et al. 2020), while efforts in low-resource languages are limited, such as Chinese (Jiang et al. 2022). To overcome this resource scarcity, cross-lingual transfer learning can be a solution. Most studies in sexism detection focus on investigating the superior model architecture for classification in different languages (Parikh et al. 2019; Rodríguez-Sánchez, Carrillo-de Albornoz, and Plaza 2020; Samory et al. 2021) without using additional domain knowledge, such as a domain-specific lexicon (Wiegand et al. 2018). Several works demonstrate the positive influence on the broader abusive language detection task by directly injecting external domain knowledge at the model level (Koufakou et al. 2020), but lack further exploration into the effect of this knowledge.

Integrating structured external knowledge like distinct lexico-semantic relations into the feature space yields better performance in various downstream tasks, such as spoken language understanding (Kim, de Marneffe, and Fosler-Lussier 2016), text simplification (Ponti et al. 2018), and cross-lingual transfer of resources (Vulić, Mrkšić, and Korhonen 2017; Ponti et al. 2019). Semantic specialisation, referred to as retrofitting or post-processing, is a process of fine-tuning pre-trained distributional word vectors by incorporating structured linguistic constraints from external lexical resources (e.g., WordNet or BabelNet) to highlight specific semantic relations in the specialised embedding space, leading to the benefit of downstream applications (Faruqui et al. 2015; Mrkšić et al. 2017). However, to overcome the restriction of constraint-driven specialisation only for existing (seen) words, a post-specialisation technique is proposed to leverage implicit information extracted from an initially specialised vector space to further specialise the entire vector space (on unseen words) (Vulić et al. 2018; Glavaš and Vulić 2018b). In addition, post-specialisation approaches are designed for cross-lingual scenarios by transferring global specialisation via a shared vector space (Glavaš et al. 2019; Ponti et al. 2019). Previous studies have developed semantic specialisation techniques for distributional word embeddings (Mrkšić et al. 2017) and contextualised embeddings with sentence-level semantics (Vulić et al. 2021). As far as we know, no previous work has studied the fine-tuning of word embeddings with domain-specific semantic knowledge through cross-lingual semantic specialisation techniques for a low-resource social media task such as sexism detection.

In this paper, we develop a domain-aware cross-lingual semantic specialisation framework between languages (i.e. English-to-Chinese), aiming to construct sexism-specific word embeddings (SexWEs) to facilitate the performance of the sexism detection task for a low-resource language. Inspired by the cross-lingual specialisation method in Ponti et al. (2019)'s work, in our case, we first structure linguistic constraints from external sexism-related semantic knowledge (e.g. BabelNet) into different forms, including source constraints (English), target constraints (Chinese) and cross-lingual constraints. Then we project all source constraints into target constraints, and refine these projected target constraints by cleaning up the noise inside them. After that, various target constraint groups are incorporated together into the specialisation process to retrofit pre-trained word embeddings to be domain-aware for the target language. Finally, we can monolingually employ our domain-specific SexWEs to the downstream task of social media sexism detection.

Furthermore, we verify the quality of our SexWEs in the intrinsic evaluation of word similarity, as well as the impact on sexism detection. Our results show that SexWEs achieves state-of-the-art performance on several word similarity benchmarks, outperforming all baseline classifiers on identifying sexism. Additionally, the visualisation of SexWEs with diverse constraints shows positive changes before and after the specialisation, and an ablation study also demonstrates the effectiveness of our proposed architecture for cross-lingual domain-aware specialisation. Our specialisation method enables us to specialise any type of distributional vectors in the target language with diverse constraints.

Our key contributions include the following:

1. We conduct the first study on semantic specialisation for cross-lingual abusive language detection, building SexWEs, sexism-specific word embeddings for Chinese;
2. Our domain-aware embeddings achieve state-of-the-art performance on word similarity benchmarks (correlation score increased by 0.039) and the Chinese sexism detection task compared with all Chinese baseline embeddings (F1 score improved by 0.114);
3. Our cross-lingual domain-aware specialisation outperforms previous state-of-the-art specialisation transfer approaches on both word similarity benchmarks (correlation score increased by 0.027) and the Chinese sexism detection task (F1 score improved by 0.041);
4. We will publicly release our resources[1] to facilitate the integration of external lexical domain knowledge into distributional embedding models for other low-resource languages.

Related Work

Sexism Detection

Research in social media sexism detection has increased in recent years (Parikh et al. 2019; Samory et al. 2021). The first attempt was by Hewitt, Tiropanis, and Bokhove (2016) who investigated the manual classification of gender-based tweets, and the first survey of automatic misogyny identification in social media was conducted by Anzovino, Fersini, and Rosso (2018). Rodríguez-Sánchez, Carrillo-de Albornoz, and Plaza (2020) explore the feasibility of automatically identifying sexist content using both traditional and deep learning techniques. In addition, researchers mainly address the problem of multilingual sexism detection by using deep neural networks with cross-Lingual word embeddings or multilingual pre-trained models (Pamungkas, Basile, and Patti 2020; Rodríguez-Sánchez et al. 2021). However, most relevant studies investigate monolingual or multilingual sexism detection only based on existing data in high-resource languages such as English and other Indo-European languages (Fersini, Nozza, and Rosso 2020; Chiril et al. 2020; Samory et al. 2021), while cross-lingual studies in the field of sexism and even general abuse are still limited for low-resource languages like Chinese.

Moreover, most studies propose model architectures for identifying online sexism or related abuse, and few make efforts to infuse external domain knowledge into vector space to enhance the detection performance (Arango, Pérez, and Poblete 2021). Badjatiya et al. (2017) utilise an LSTM-based model to generate English hate word embeddings, but more persuasive validation strategies should be reconsidered (Arango, Pérez, and Poblete 2020). Kamble and Joshi (2018) describe the construction of domain word embeddings based on Word2Vec from a Hindi-English hate speech dataset, and Alatawi, Alhothali, and Moria (2021) produce abuse-specific embeddings for English white supremacy. Besides, multilingual word embeddings based on abuse knowledge are created for cross-lingual hate speech detection (Arango, Pérez, and Poblete 2021).

To the best of our knowledge, no prior work has studied cross-lingual semantic specialisation techniques to generate domain-aware word embeddings in low-resource languages for abuse in social media. Therefore, we specifically apply this technique to the field of sexism and choose Chinese[2] as our target language. Abusive language detection is in turn

[1] https://github.com/aggiejiang/SexWEs

[2] Chinese is generally a resourceful language, but there is only one dataset available for online abuse or sexism.

understudied in Chinese, with the only antecedent of (Jiang et al. 2022), who created the sexism dataset that we use here but didn't study model development.

Retrofitting Word Embeddings

In the field of word vector specialisation, there has been a body of research exploring various methods to incorporate diverse constraints into the word embedding space. The first retrofitting work by Faruqui et al. (2015) is proposed to pull the vectors of similar words closer to each other by fusing only synonyms. Then ATTRACT-REPEL, a standard semantic specialisation approach, is developed to integrate structured linguistic constraints with both similar and dissimilar semantics into pre-trained vector spaces, clustering the embeddings of similar words (e.g. synonyms, hypernym-hyponym pairs) closer together and enforcing dissimilar words (e.g., antonyms) far away from each other (Mrkšić et al. 2017). Such semantic specialisation could be applied to any kind of distributional word embeddings.

Since the first-generation semantic specialisation models only retrofit the embeddings of words seen in linguistic constraints, a series of post-specialisation techniques are proposed (Vulić et al. 2018; Glavaš and Vulić 2018b; Ponti et al. 2018; Colon-Hernandez et al. 2021). Post-specialisation aims to fine-tune the entire distributional vector space by learning an explicit and global specialisation mapping between original and initially specialised spaces, and then applying the mapping to the embeddings of words unseen in external constraints (Vulić et al. 2018). Ponti et al. (2018); Colon-Hernandez et al. (2021) modify the feed-forward post-specialisation network with different Generative Adversarial Networks (GANs) based approaches to discriminate word vectors from original and specialised spaces, which yields better performance on retrofitting.

Post-specialisation approaches can be further employed for cross-lingual transfer through a shared vector space between source and target languages (Glavaš et al. 2019; Ponti et al. 2019). In this work, we demonstrate how to combine task-oriented multilingual domain knowledge to achieve cross-lingual semantic specialisation on pre-trained word embeddings, with an impact on sexism detection for low-resource languages.

Methodology: SexWEs

We propose to build Sexist Word Embeddings (SexWEs) based on a cross-lingual domain-aware semantic specialisation system, inspired by the CLSRI framework Ponti et al. (2019). The objective is to incorporate awareness of the sexism domain into the semantic specialisation procedure to enrich domain-aware word embeddings (integrated sexism domain knowledge). We aim to specialise existing state-of-the-art distributional word embeddings in a target language by utilising commonsense knowledge and multilingual domain knowledge from lexical constraints, where constraints are dominated by resource-rich source language and supplemented by a resource-poor target language. In our case, we opt for English (EN) as the source language L_{en} and Chinese (ZH) as the target language L_{zh}.

Our procedure can be split into two parts: constraint processing and domain-aware specialisation (see Figure 1). Firstly, constraint processing is to collect multilingual domain constraints, project source constraints across languages and clean up noisy constraints by transformation. Then we fuse the refined target constraints and external target constraints together as constraints C_{zh}^{group}, and execute monolingually initial specialisation and post-specialisation on existing distributional word vector space by employing well-handled constraints C_{zh}^{group} in the target language.

Constraint Processing

According to Mrkšić et al. (2017)'s ATTRACT-REPEL methodology, linguistic constraints obtained from external sources are usually divided into two lexico-semantic groups:

- ATTRACT *constraints*: indicate word pairs with similar representations, e.g. synonyms (swearing and abuse, 咒骂and 辱骂) or direct hypernym-hyponym pairs (woman and widow, 女人and 寡妇);
- REPEL *constraints*: specify which word pairs should appear far apart in the vector space, e.g. antonyms (appreciation and disgust, 欣赏and 厌恶).

Our constraints are grouped into five categories:

- English general constraints C_{en}^{g}
- English domain constraints C_{en}^{d}
- English general&domain constraints $C_{en}^{both} = C_{en}^{g} \cup C_{en}^{d}$
- Chinese domain constraints C_{zh}^{d}
- Cross-lingual EN-ZH domain constraints C_{cl}^{d}

English general constraints include words that are commonly and frequently used, while domain constraints refer to words related to the domain. In our case, we continue to use the existing general constraints (Ponti et al. 2019) and extract domain constraints in both monolingual and cross-lingual scenarios. Except for C_{cl}^{d} constraints, the other four types of constraints all have ATTRACT and REPEL sets separately. This step focuses on processing source constraints and cross-lingual constraints while target constraints C_{zh}^{g}, C_{zh}^{d} or C_{zh}^{both} could be regarded as external constraints to facilitate specialisation performance in the next step.

In this constraint processing step, we first collect domain constraints into ATTRACT and REPEL sets from BabelNet in source language C_{en}^{d} and target language C_{zh}^{d}, and project all constraints in source language C_{en}^{d} to those in target language ${C_{zh}^{d}}'$. Considering imperfect mapping and polysemy of C_{en}^{d} possibly leading to the incorrect meaning of ${C_{zh}^{d}}'$, these noisy constraints ${C_{zh}^{d}}'$ are filtered via a variant of Specialisation Tensor Model (STM) (Glavaš and Vulić 2018a).

Monolingual and Cross-lingual Domain Constraints Collection In order to extract monolingual domain constraints, we organise domain seed words from several domain-related lexical resources for both source C_{en}^{d} and target C_{zh}^{d} languages separately. Then we create domain constraint pairs via searching synonyms and antonyms in the

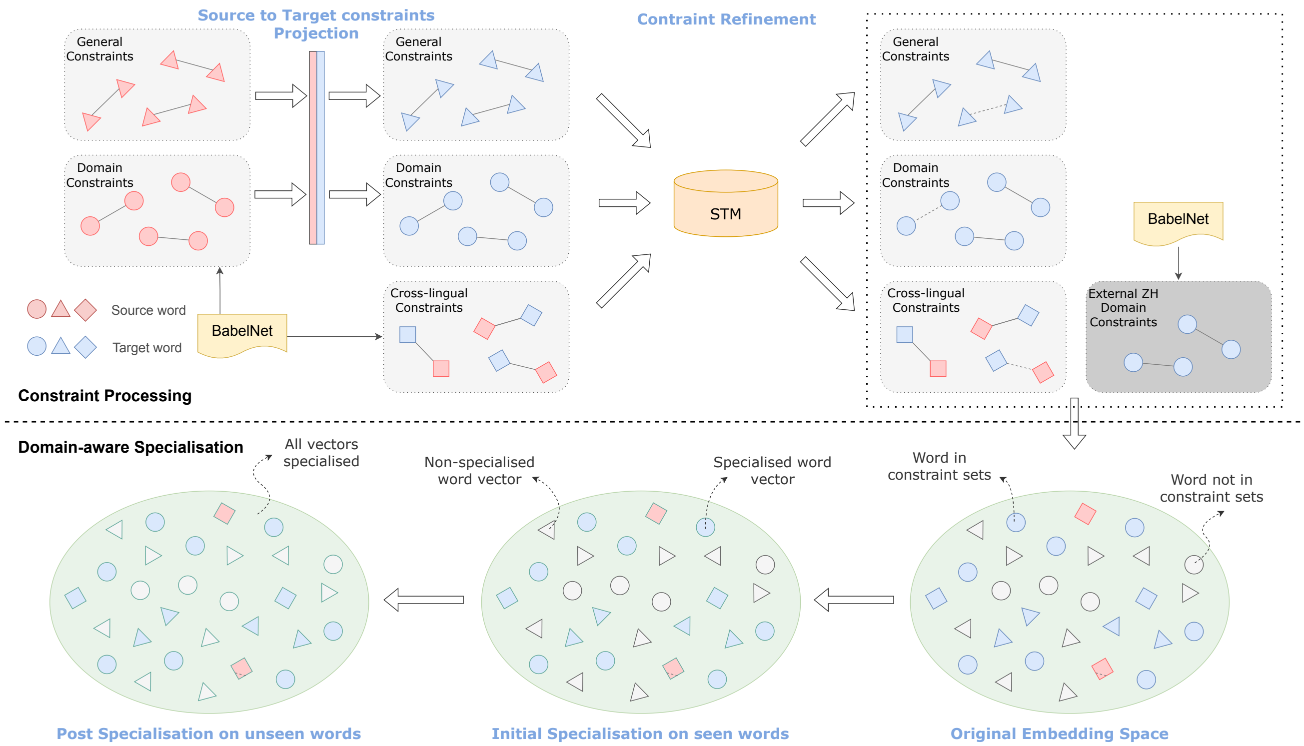

Figure 1: Overview of SexWEs. Constraint processing collects multilingual domain constraints, projects them across languages and filters noisy pairs. Domain-aware specialisation retrofits distributional word vectors in two steps: (1) utilise knowledge-aware constraints to specialise vectors on seen words; (2) learn and apply specialised mapping to the entire space.

same language for each seed word, and add a language tag before each word, such as (zh_歧视, zh_偏见)[3]

In addition to monolingual domain constraints, we also extract cross-lingual domain constraints C^d_{cl} based on domain seed words in the form of English-Chinese constraints. It will be taken into consideration such as explicit and implicit cross-lingual domain constraints. An explicit constraint refers to those English-Chinese constraints via direct translation and both words are explicitly domain-related (such as (en_prejudice, zh_歧视)), while an implicit constraint means two words cannot be directly translated from/to each other, because one word is domain-related in one language but another one could be domain-unrelated in another language if directly translated (such as (en_f*cking, zh_草)[4] and (zh_绿茶婊, en_angelic b*tch)[5]). All domain seed words are first directly translated[6] into explicit constraints, then we manually check and correct incorrectly translated word pairs to generate implicit constraints.

[3]歧视or 偏见means an unfair and unreasonable opinion or feeling, especially when formed without enough thought or knowledge, such as prejudice or bias.

[4]The primary meaning of 草is grass, only in certain occasions it may mean the same as f*cking.

[5]绿茶婊refers to girls who pretend to be pure and innocent but in fact are manipulative and scheming. It literally translates into green tea b*tch. The meaning of 绿茶婊is similar to angelic b*tch

[6]We use Google Translate https://translate.google.co.uk/.

Source to Target Constraints Projection Learning cross-lingual word embeddings via supervised approaches shows good performance on the task of Bilingual Lexicon Induction (BLI) especially on typologically-distant language pairs like EN-ZH (Wang, Henderson, and Merlo 2021). Recent work (Wang, Henderson, and Merlo 2021) has shown that Relaxed Cross-domain Similarity Local Scaling (RCSLS) (Joulin et al. 2018), as a supervised system, achieves remarkable performance among competing models on the BLI task, and it has been applied to the word translation task in order to enhance the performance. So we leverage the RCSLS model to learn a linear cross-lingual projection matrix $\mathbf{W}_{en\_zh}$ between source and target word embeddings.

Given a set of source constraints C_{en}, each constraint is presented as a word pair (w^a_{en}, w^b_{en}). Since phrases exist widely in domain constraints C^d_{en}, phrase-level projection is also employed by averaging all word embeddings per phrase. We translate each word or phrase w_{en} in source constraints by looking for the nearest neighbour of its (averaged) embedding $\mathbf{x}_s$ in the projected target space. We project source constraints C_{en} into target constraints $C^d_{zh'}$ by using the projection matrix $\mathbf{W}_{en\_zh}$ to project source and target embeddings into a shared bilingual space $\mathbf{X}_{en\_zh}$.

Target Constraint Refinement The shared bilingual space obtained by the cross-lingual projection matrix is far from perfect due to incorrect translation via the cross-lingual shared space and incorrect of senses of polysemous words in L_{en}. Hence, noisy constraints could be generated via projection-based approaches from source constraints C_{en} to target constraints $C^{d}_{zh'}$ (Glavaš et al. 2019).

Similar to the CLSRI framework, we aim to purify noisy constraints in $C^{d}_{zh'}$ by leveraging the Specialisation Tensor Model (STM) to discriminate lexico-semantic relations within word pairs (Glavaš and Vulić 2018a). STM is a simple and effective feed-forward neural architecture that predicts lexical relations between word pairs by specialising input distributional word embeddings in multiple different projections and computing latent scores from these specialisation tensors for the final relation classifier. STM performs better particularly for synonyms and antonyms, and also presents stable performance across languages (Glavaš and Vulić 2018a). We alter the multi-label STM classifier to a binary classifier[7], and train five types of instances for STM:

- G_a-STM & D_a-STM: it predicts whether a word pair from general or domain constraints represents a valid ATTRACT constraint;
- G_r-STM & D_r-STM: it predicts whether a word pair from general or domain constraints represents a valid REPEL constraint;
- D_{cl}-STM: it predicts whether a pair of cross-lingual domain words represents a valid ATTRACT constraint;

Domain-Aware Specialisation

The step of domain-aware specialisation consists of monolingually retrofitting distributional word embeddings space in the target language L_{zh} by leveraging a group of target constraints，such as projected target constraints (e.g. $C^{g}_{zh'}$, $C^{d}_{zh'}$ or $C^{both}_{zh'}$) plus external target constraints (e.g. C^{g}_{zh}, C^{d}_{zh} or C^{both}_{zh}). The whole semantic specialisation process is similar to the CLSRI system (Ponti et al. 2019). Following the state-of-the-art specialisation model ATTRACT-REPEL (AR) (Mrkšić et al. 2017), we initially specialise the target distributional space to be domain-aware but limited to existing C_{zh} constraints. Then, based on the AR specialisation, we apply the state-of-the-art post-specialisation model RETROGAN (Colon-Hernandez et al. 2021) to the entire vocabulary V_{zh}, including all the words seen and unseen in the target space. The following is a detailed description of system and a brief outline of AR and RETROGAN models.

Initial Domain-Aware Specialisation The group of target constraints C^{group}_{zh} to be specialised is a combination of projected target constraints $C_{zh'}$ from source constraints C_{en} and external target constraints C_{zh} from scratch, where $C^{group}_{zh} = C^{type}_{zh} \cup C^{type}_{zh'}$ and $type = \{g, d, both\}$. After the combination, C^{group}_{zh} includes two constraint subsets: ATTRACT constraints A_{zh} and REPEL constraints R_{zh}. The distance of each word pair (w^{a}_{zh}, w^{b}_{zh}) from A_{zh} and R_{zh} is refined between their corresponding embeddings $(\mathbf{x}^{a}_{zh}, \mathbf{x}^{b}_{zh})$ in the target distributional space.

The specialisation process is carried out via mini-batches of C^{group}_{zh}. Let $\mathcal{B}_A$ be a batch of vector pairs from A_{zh} and $\mathcal{B}_R$ the batch from R_{zh}. We define $\mathcal{T}_A(\mathcal{B}_A)$ and $\mathcal{T}_R(\mathcal{B}_R)$ as corresponding negative pairs for each $\mathcal{B}_A$ and $\mathcal{B}_R$. For each A_{zh} (or R_{zh}) constraint $(\mathbf{x}^{a}_{zh}, \mathbf{x}^{b}_{zh})$, we retrieve its closest (or farthest) vector pair as the negative constraint $(\mathbf{t}^{a}_{zh}, \mathbf{t}^{b}_{zh})$. Half of the negative constraints are selected based on their cosine similarity, and the other half are random negative samples.

The objective of AR retrofitting is to minimise the max margin loss between target constraints and their corresponding negative samples, which includes three types of losses:

$$\mathcal{L}_{AR} = Att(\mathcal{B}_A, \mathcal{T}_A) + Rep(\mathcal{B}_R, \mathcal{T}_R) + Pre(\mathcal{B}_A, \mathcal{B}_R) \quad (1)$$

Specifically, $Att(\mathcal{B}_A, \mathcal{T}_A)$ enables target constraints in $\mathcal{B}_A$ closer together than those in the corresponding $\mathcal{T}_A$ by a ATTRACT margin δ_A:

$$\begin{aligned} Att(\mathcal{B}_A, \mathcal{T}_A) = \sum_{i=1}^{|\mathcal{B}_A|} [&\mathcal{T}(\delta_A + \mathbf{x}^{a}_{zh_i}\mathbf{t}^{a}_{zh_i} - \mathbf{x}^{a}_{zh_i}\mathbf{x}^{b}_{zh_i}) \\ &+ \mathcal{T}(\delta_A + \mathbf{x}^{b}_{zh_i}\mathbf{t}^{b}_{zh_i} - \mathbf{x}^{a}_{zh_i}\mathbf{x}^{b}_{zh_i})] \end{aligned} \quad (2)$$

where $\mathcal{T}(x) = max(0, x)$ is the hinge loss function, and δ_A determines how much closer target constraints from A_{zh} are to each other than the distance to their corresponding negative examples. Analogously, $Rep(\mathcal{B}_R, \mathcal{T}_R)$ imposes constraints in $\mathcal{B}_R$ farther than their corresponding constraints in $\mathcal{T}_R$ based on a REPEL margin δ_R. Besides, $Pre(\mathcal{B}_A, \mathcal{B}_R)$ is the regularisation term to preserve the high-quality semantic information from $\mathbf{X}_{zh}$ by minimising the Euclidean distance between plain and specialised embeddings.

After AR specialisation, AR specialised space $\mathbf{X}^{'}_{zh} \in \mathcal{R}^d$ is generated from the initial distributional space $\mathbf{X}_{zh} \in \mathcal{R}^d$.

Cyclic Adversarial Post-Specialisation The AR specialisation only works on the target words V^{seen}_{zh} that actually exist in C^{group}_{zh}, which indicates that the performance of initial specialisation can be semantically improved in terms of the overlapping vocabulary between explicit V^{seen}_{zh} and the vocabulary V_{zh} of the initial distributional space $\mathbf{X}_{zh}$. Post-specialisation learns the mapping from initial specialisation space and propagates it to the rest of the vocabulary V^{unseen}_{zh} (Vulić et al. 2018; Colon-Hernandez et al. 2021).

RETROGAN enriches the existing adversarial post-specialisation model (Ponti et al. 2018) to a CycleGAN-like architecture with a pair of Generative Adversarial Networks (GANs) (Goodfellow et al. 2020). The goal of RETROGAN is to learn a global specialisation mapping by balancing a combination of losses in both post-specialisation and inversion to ensure a unique one-to-one mapping between the plain vector space $\mathbf{X}_{zh}$ and specialised AR space $\mathbf{X}^{'}_{zh}$ as conditioned by embeddings of seen words V^{seen}_{zh} from C^{group}_{zh} constraints. Then it propagates this global mapping to the entire distributional space of our target language $\mathbf{X}_{zh}$.

The model combines both cyclic and non-cyclic optimisation objectives, where the contrastive margin-based ranking loss with random confounders L_{MM} (Ponti et al. 2018, 2019) is used for both the generators and additionally for the

[7]See more STM technical details in Glavaš and Vulić (2018a)

cycle of generators[8]:

$$\mathcal{L}_{MM} = \sum_{i=1}^{||V_{zh}^{seen}||} \sum_{j=1|j\neq i}^{k} \mathcal{T}$$
$$[(\delta_{MM} - \cos(G(\mathbf{x}_{zh_i}), \mathbf{x}'_{zh_i}) + \cos(G(\mathbf{x}_{zh_i}), \mathbf{x}'_{zh_j}) +$$
$$(\delta_{MM} - \cos(F(\mathbf{x}_{zh_i}), \mathbf{x}'_{zh_i}) + \cos(F(\mathbf{x}_{zh_i}), \mathbf{x}'_{zh_j}) +$$
$$(\delta_{MM} - \cos(G(F(\mathbf{x}_{zh_i})), \mathbf{x}'_{zh_i}) + \cos(G(F(\mathbf{x}_{zh_i})), \mathbf{x}'_{zh_j}) +$$
$$(\delta_{MM} - \cos(F(G(\mathbf{x}_{zh_i})), \mathbf{x}'_{zh_i}) + \cos(F(G(\mathbf{x}_{zh_i}), \mathbf{x}'_{zh_j}))] \quad (3)$$

where $G : \mathbf{X}_{zh} \rightarrow \mathbf{X}'_{zh}$ is the generator that maps the plain vector space $\mathbf{X}_{zh}$ to the specialised space $\mathbf{X}'_{zh}$, and $F : \mathbf{X}'_{zh} \rightarrow \mathbf{X}_{zh}$ is the generator that does the opposite. L_{MM} makes a word vector generated from $\mathbf{X}_{zh}$ by generators closer to its gold-standard vector (e.g. specialised AR vector $\mathbf{x}'_{zh} \in \mathbf{X}'$) and different from any of k random confounders by a margin δ_{MM}, and then forces this constraint across the cycle.

Experimental Setup

Initial Distributional Word Embeddings

As a starting point to build domain-aware specialised embeddings, we employ publicly available FASTTEXT word vectors (Grave et al. 2018) for both English and Chinese.[9] They provide 300-dimensional word vectors trained on Common Crawl and Wikipedia in 157 languages, using CBOW with position weights. We execute the projection from source to target vector space via supervised RCSLS method, searching 10 nearest neighbours in 10 iterations.

External Sexism Lexical Knowledge

To generate domain-specific constraints, we intend to use some lexical resources related to sexist domains to organise sexist seed words. However, due to the lack of external resources specifically addressing sexism, we select words from abusive language-related resources, where abuse is a superdomain of sexism (Waseem and Hovy 2016).

For the source language (EN), we use (i) the hate speech lexicon *HurtLex*, containing 6,287 seed offensive, aggressive, and hateful words and phrases in over 50 languages (Bassignana, Basile, and Patti 2018), and (ii) the abuse lexicon by *Wiegand et al. (2018)*, which includes 2,989 words. For the target language (ZH), we use *SexHateLex* (Jiang et al. 2022), a large Chinese sexism lexicon including 3,016 profane and sexually abusive and slang words and phrases.

Linguistic Constraints

Linguistic Constraints are present in the form of word/phrase pairs in the source language (EN) and the target language (ZH) for semantic specialisation, which is divided into three categories: source general constraints, bilingual domain (sexism) constraints and cross-lingual constraints. We also combine general and domain constraints (in the same language) as another group of constraints. The number of constraints is summarised in Table 1.

- *Source General Constraints:* We follow the same English general constraints as used in previous work for the specialisation process (Ponti et al. 2018, 2019). These general constraints involve the lexico-semantic relations from WordNet (Miller 1995), Paraphrase Database (PPDB) (Ganitkevitch, Van Durme, and Callison-Burch 2013) and BabelNet (Navigli and Ponzetto 2010), which covers 16.7% of the 200K most frequent English words in the vocabulary of FASTTEXT embeddings.
- *Bilingual Domain Constraints:* To produce domain constraints, we employ the multilingual semantic network BabelNet on sexism-related seed words or phrases to extract synonyms and antonyms according to word sense tags. These constraints cover only 14.4% and 4.2% of the English and Chinese vocabulary from FASTTEXT.
- *Cross-lingual Domain Constraints:* Cross-lingual sexism-related (domain) constraints are English-Chinese pairs extracted via multilingual BabelNet based on domain seed words or phrases (e.g. en_hate, zh_憎恶).

		General	**Sexism**	**Both**
English	ATTRACT	640,435	130,445	768,294
	REPEL	11,939	501	12,148
Chinese	ATTRACT	-	6,353	-
	REPEL	-	32	-
EN-ZH	ATTRACT	-	189	-

Table 1: Collection of ATTRACT and REPEL constraints for source (EN) and target (ZH). Both is the aggregate and deduplicated set of general and sexism-related constraints.

Specialisation Approaches in Comparison

We compare our SEXWES specialisation on different types of constraints with three other semantic specialisation methods, implemented using the same FASTTEXT embeddings and both constraints used for our model SEXWES:

- ATTRACT-REPEL *(*AR*):* A state-of-the-art retrofitting approach (Mrkšić et al. 2017) to refine a distributional vector space by using ATTRACT/synonymy and REPEL/antonymy constraints.
- RETROGAN : A post-specialisation approach (Colon-Hernandez et al. 2021) by learning the mapping of AR and then extending an adversarial post-specialisation model AUXGAN (Ponti et al. 2018) into a CycleGAN-like architecture (Zhu et al. 2017) on the entire dataset.
- CLSRI *:* A specialisation Transfer via Lexical Relation Induction (Ponti et al. 2019) transfers specialisation mapping from a resource-rich source language (English) to virtually any target language based on AR and AUXGAN with noisy constraints cleanup.

[8]See more technical details of the RETROGAN and its losses in Colon-Hernandez et al. (2021)

[9]Other multilingual embedding models, such as LASER, Multilingual BERT and XLM-R, could be tested, however they are generally better suited for sentence-level embeddings.

Hyperparameters in the Training Process

Constraints Refinement: STM The STM model is adopted to predict lexical relations between constraints with 5 specialisation tensors, 300 neurons of the hidden layer and a 0.5 dropout value based on prior work (Ponti et al. 2019). During training, we set the batch size to 32 and the maximum number of iterations to 10, using Adam optimiser (Kingma and Ba 2015) with a learning rate of 0.0001.

Initial specialisation: AR We preserve the hyperparameter settings for AR as used by Mrkšić et al. (2017). The margins for ATTRACT, REPEL and regularisation are 0.6, 0.0 and $1e^{-9}$, respectively. The Adagrad optimiser (Duchi, Hazan, and Singer 2011) is used with 0.05 learning rate, batch size is 50, and maximum number of iterations is 5. The same configuration as the baseline AR.

Post-Specialisation: RETROGAN We use two hidden layers with 2,048 units for the generator and the discriminator in each GAN of RETROGAN , adopting 0.2 and 0.3 dropout rates separately. We set the margin δ_{MM} to 1.0 and the number of negative samples to 25, utilising Adam optimiser with 0.1 learning rate. The number of training epochs is set to 10 and batch size 32, same as the baseline RETROGAN model.

Results and Analysis

We evaluate our SexWEs via both intrinsic evaluation of word similarity and extrinsic evaluation of sexism detection.

Intrinsic Evaluation : Word Similarity

The first experiment is to assess the quality of our specialised space of SexWEs via the word similarity task, which aims to evaluate the ability of the model to capture the semantic proximity and relatedness between two words.

Chinese Embeddings in Comparison We adopt original FASTTEXT word vectors and retrofitted vectors by other specialisation approaches in comparison with our specialised embeddings infusing diverse constraints.

Evaluation Setup We employ three word similarity benchmarks, namely SimLex-999 (SL999) (Hill, Reichart, and Korhonen 2015), WordSim-296 (WS296) (Jin and Wu 2012) and WordSim-240 (WS240) (Wang et al. 2011). WS296 and WS240 are Chinese datasets, while SL999 is an English dataset then translated into traditional Chinese by Su and Lee (2017). We convert it from traditional to simplified Chinese with chinese-converter[10]. The word pair coverage in the datasets is 975 of 999 for SL999, 230 of 240 for WS240, and 286 of 297 for WS296. The Spearman's rank correlation ρ is measured as the intrinsic evaluation metric between the gold word pair similarity scores by annotators and the cosine similarity scores of the corresponding word embeddings from various vector spaces.

Analysis of Results The results of word similarity tests are summarised in Table 2. Regardless of whether we plus external Chinese domain constraints or not, our specialised SexWEs basically outperforms the initial distributional vectors (0.039) and other cross-lingual specialisation models (0.027), indicating the effectiveness of incorporating domain constraints in source language during the cross-lingual transfer. And to the best of our knowledge, our results also surpass the Chinese word embeddings VCWE (Sun, Qiu, and Huang 2019) that achieves the state-of-the-art performances on WS240 and WS296[11]. By fusing external domain-specific target pairs, it also achieves better results for vector space specialisation. Moreover, even without the infusion of sexist-related knowledge, our approach still outperforms the similarly structured model CLSRI , while noticeably exceeding two separate models of AR and RETROGAN, respectively. Although our SexWEs achieves a satisfactory performance on SL999 among all models, it can still be noted that there is no big gap compared to scores on the other two benchmarks, probably due to the translation issue from English to Chinese version or the conversion issue between traditional and simplified Chinese.

	SL999	WS240	WS296
FASTTEXT	.347	.546	.620
AR	<u>.402</u>	.521	.586
RETROGAN	.380	.572	.615
CLSRI	.384	.558	<u>.627</u>
SexWEs	**.406**	**.586**	.608
w/o external	.394	<u>.581</u>	.624
only general	.389	.561	.623
only domain	.388	.563	**.637**

Table 2: Results of word similarity evaluation based on Spearman's rank correlation score ρ (average of 5 runs).

Extrinsic Evaluation: Sexism Detection

We next implement extrinsic evaluation to adjust our specialised SexWEs to a downstream binary classification task – sexism detection – which assesses the effectiveness of word embeddings with domain information.

Dataset We use the only sexism dataset in Chinese, Sina Weibo Sexism Review (SWSR) (Jiang et al. 2022), with posts labeled for sexism from the Sina Weibo platform. SWSR annotations are constructed at different levels of granularity, and we use the binary labels: sexist and non-sexist. We split the entire dataset into training and test sets in the ratio of 4 to 1. We further randomly select 20% of the training set as the validation set for model fine-tuning process, and finally utilise the whole training set to evaluate model capacity on test set. More details are shown in Table 3.

Sexism Detection Models Tested We leverage a simple text-based Convolutional Neural Network (TCNN) (Kim 2014) as our primary classifier, which is a popular architecture for dealing with NLP tasks with a good feature extraction capability (Zhang and You 2021), leading to a smaller

[10]https://pypi.org/project/chinese-converter/

[11]The VCWE results are 0.578 for WS240 and 0.613 for WS296, and it exceeds many competitive Chinese embeddings (Xiong et al. 2021). For more results, see https://chinesenlp.xyz/docs/word_embedding.html.

	Train	Validation	Test
Sexist	2244	561	288
Non-Sexist	4214	1053	609
Total	6458	1614	897
STR (%)	34.7	34.8	32.1

Table 3: Distribution of train, validation and test sets, sexist text rate (STR) in the SWSR dataset.

number of parameters, lower computational needs, and a faster training speed (Zhang and You 2021). TCNN is fed with different vectors used in the intrinsic evaluation, or changed to other state-of-the-art models for comparison to demonstrate the impact of our specialised embeddings on detecting sexist text. For vectors, we use the original and specialised word embeddings evaluated in the intrinsic experiments in combination with static BERT embeddings extracted from Chinese BERT[12].

As baseline models, we use BERT (Devlin et al. 2019) and a state-of-the-art Chinese pre-trained model MACBERT[13], which performs better than normal Chinese BERT[14] and other variants in some classification tasks (Cui et al. 2020).

Evaluation Setup We use the Adam optimiser (0.0001 learning rate) and a maximum sequence length of 100 for all baseline models. TCNN contains 128 units in the hidden layer with the dropout value 0.4, and we use Huggingface models 'bert-base-chinese' (BERT) and 'hfl/chinese-macbert-base' (MACBERT). We train the TCNN-based models for 100 epochs and BERT-based models for 4 epochs, using the same batch size of 32. We report the accuracy and macro F1 scores as the evaluation metrics.

Analysis of Results We report the results for sexism detection in Table 4. We see that the classifier with our SexWEs achieves the highest F1 and accuracy scores, outperforming all baseline classifiers and classifiers with baseline retrofitted embeddings, and most of our models with different constraints display better results than baselines. The classifier with our SexWEs also exhibits stable performance with relatively small fluctuations in scores. Comparing baseline embeddings, there are notable improvements (0.093-0.135) in our SexWEs compared to those using FASTTEXT word embeddings and popular Chinese embeddings VCWE, and better performance than BERT embeddings. Additionally, our model slightly outperforms BERT-related models BERT and MACBERT, but both of them present smaller fluctuations due to high stability. RETROGAN shows the best results among all baseline specialisation models and outperforms all non-specialised embeddings, but it is still below our SexWEs. Moreover, we can draw some conclusions that are in line with the intrinsic evaluation. That is, leveraging sexism-related constraints and external constraints in target language for the cross-lingual specialisation process improves the detection of online sexism, and only using general constraints also shows the effectiveness in this task compared to other specialisation baselines.

Impact of Class Imbalance Sexism or abuse tends to be the minority class in most datasets. In the case of the SWSR dataset, 65.5% are non-sexist instances (Jiang et al. 2022). Our SexWEs F1 score for the sexist class is 0.626, which is still clearly below the F1-not score of 0.849. We can also clearly observe that the F1-scores between sexist and non-sexist classes differ greatly, with the average F1-score of the non-sexist class being about 0.227 higher than that of the sexist class. This shows a negative impact of class imbalance on the sexism detection task, and the potential challenges that sexist texts may bring to the detection (see more in the subsection Qualitative Analysis).

Resampling and data augmentation techniques could be considered to mitigate the imbalance in the future (Bigoulaeva et al. 2022; Rizos, Hemker, and Schuller 2019).

Qualitative Analysis In addition to quantitative evaluation, we also conduct a qualitative analysis of some misclassified cases to assess the potential of SexWEs for sexism detection as well as the challenges. When looking at predicted examples from BERT and classifiers with original embeddings and our SexWEs, we see some recurrent types of misclassification in Table 5.

(i) Implicit sexism: Humour, irony and sarcasm are difficult to be identified. Example (1) is sexist irony without an explicitly abusive expression. The model with SexWEs successfully deemed it sexist, while the others failed.

(ii) Informal and code-mixed expressions: Example (1) is a Chinese-English code-mixed text, and the slang word 'vans' in English has a similar pronunciation as '完事了' (that's it) in Chinese. '驴' usually refers to 'donkey', but it is commonly used in sexist expressions that offend women[15].

(iii) Implicit attack target: The attack target might not explicit appear like example (2). All models failed to predict it as sexist text. It demeans the group of highly educated males, but the target can only be guessed from the context.

(iv) Homophones: Homophones are common in sexist speech to convey abusive connotations, or to obfuscate and avoid detection. '说服' and '睡服' have the same pronunciation in (2). '说服' is a general term that means persuade or convince, and '睡服' is a homophonic word with a similar meaning to persuade someone by f*cking.

(v) Overuse of explicit sexist words: Sexist words might be overused in one text, leading to the over-dependence of the model on these words, while sexist targets in posts are confounding and hard to be identified. All models failed in example (3), and we see that the model can easily deem a text sexist if it contains many sexist words, despite not having a specific targeted individual or group.

[12]We extract contextualised BERT embeddings from the initial embedding layer of Chinese BERT trained on SWSR training set, using Huggingface BERT model 'hfl/chinese-bert-wwm-ext'.

[13]https://huggingface.co/hfl/chinese-macbert-base

[14]https://huggingface.co/bert-base-chinese

[15]'驴' comes from '婚驴' (marriage donkey), and is intended to depict the image of 'women who are as stupid as donkeys in marriage, deprived of a lot of benefits, but still enjoy silly happiness'.

Model		F1-sex	F1-not	Macro-F1	Accuracy
Baseline embeddings	+FT	.483 (±.015)	.723 (±.044)	.603 (±.028)	.641 (±.040)
	+VCWE	.355 (±.149)	.796 (±.010)	.645 (±.071)	.682 (±.008)
	+BERT_emb	.573 (±.059)	.835 (±.009)	.704 (±.027)	.765 (±.006)
	+AR	.490 (±.025)	.840 (±.017)	.668 (±.009)	.770 (±.011)
	+RETROGAN	.622 (±.010)	.811 (±.056)	.717 (±.027)	.753 (±.044)
	+CLSRI	.638 (±.005)	.775 (±.010)	.707 (±.006)	.723 (±.007)
Baseline models	BERT	.641 (±.006)	.782 (±.008)	.711 (±.006)	.729 (±.007)
	MACBERT	**.658 (±.013)**	.789 (±.015)	.724 (±.013)	.739 (±.014)
SexWEs	**SexWEs**	.626 (±.035)	**.849 (±.008)**	**.738 (±.016)**	**.786 (±.008)**
	w/o external	.627 (±.041)	.840 (±.044)	**.738 (±.024)**	.761 (±.034)
	only general	.622 (±.061)	.842 (±.011)	.732 (±.030)	.779 (±.012)
	only domain	.646(±.011)	.817 (±.056)	.733 (±.032)	.764 (±.046)

Table 4: Results of sexism detection with standard deviations (average of 10 runs).

Text	TCNN+FT	BERT	SexWEs	Ground Truth
1. 尊重驴不带套的权力，意外怀孕的权力，尊重就vans。 `Translation`: Respect the rights of women without wearing condoms and unintended pregnancies, that's it.	Non-Sexist	Non-Sexist	Sexist	**Sexist**
2. 学历高的估计更厉害，从道理上说服你，不然就身体上睡服你。 `Translation`: Males with higher education may be better at persuading you or f*cking you.	Non-Sexist	Non-Sexist	Non-Sexist	**Sexist**
3. 田园女权，女拳师，极端女权，是我是我都是我。 `Translation`: Pastoral feminist, female boxer, extreme feminist, it is all me.	Sexist	Sexist	Sexist	**Non-Sexist**

Table 5: Misclassified examples by three models: TCNN + FASTTEXT embeddings (TCNN+FT), BERT, and TCNN + specialised embeddings (SexWEs), along with ground truth labels.

Discussion

Visualisation of Word Embeddings

We visualise both original FASTTEXT embeddings and various specialised SexWEs embeddings. We select six sexism-related seed words and gather each seed word with its 20 nearest neighbors from the initial word vector space, to explore changes in these domain word groups during our specialisation process. Figure 2 shows the visualisation of word embeddings with dimensional reduction by t-SNE (van der Maaten and Hinton 2008). To further investigate the semantic shift between different word vector spaces (Hamilton, Leskovec, and Jurafsky 2016), we measure the average cosine distance between a seed word and its neighbours in each local word cluster, and average distances among the six clusters to obtain the overall distance in the space. The local distance is presented in subplot titles of Figure 2.

Looking at both the spatial range of visualised word clusters and local distances, we can observe that all specialised groups of domain words become more independent and get closer from the original distributional vector space in Figure 2 (a) to any of our specialised vector space (see Figure 2 (b)-(f)), which illustrates the benefit of our specialisation method. After the specialisation process with English constraints, the distance of word clusters shows a significant decrease, further decreasing after adding external Chinese constraints. For word embeddings that incorporate more domain information (Figure 2 (e) and (f)), the connections between words in each cluster become stronger, compared to embedding spaces that are only retrofitted with knowledge of general constraints in Figure 2 (d). Furthermore, after adding external Chinese constraints, the vector space specialised only with domain knowledge becomes more contiguous (see Figure 2 (e) to (b)), while the spaces specialised by both constraints are relatively sparse (see Figure 2 (f) to (c)). This opposite change may be caused by perturbations of commonsense knowledge, since general constraints outnumber domain constraints.

Ablation Study

To evaluate different components, we perform a study of the following ablated models of SexWEs: *(i) removing phrase-level projection; (ii) removing constraint refinement;* and *(iii) removing* RETROGAN *post-specialisation*. In Table 6, we can see that our model outperforms all ablated models, which demonstrates the important contribution of all components. Although phrase-level constraint processing in the projection step does not significantly improve the quality of embeddings, this step validates the positive impact of doing domain-related phrase mapping on identifying sexism. The results also highlight the effectiveness of STM in refining the noisy lexico-semantic relations between constraints compared with the one without constraint refinement. Furthermore, we can validate the capability of RETROGAN post-specialisation step to efficiently apply the retrofitting map-

Figure 2: t-SNE visualisations of SexWEs word embeddings. Each color group indicates a Chinese domain word with its 20 neighbours generated from original FASTTEXT vectors. There are totally 6 seed words selected, namely purple for 女人(woman), blue for 性侵(sexual assault), skyblue for 强奸(rape), green for 下贱(b*tchy), orange for 傻(stupid), and red for 责骂(scold). Averaged local distance of word clusters (local_dist) is measured based on the t-SNE space.

ping to full word vector space when comparing to specialised word vectors without post-specialisation step.

	Intrinsic			Extrinsic	
	SL999	**WS240**	**WS296**	**Macro-F1**	**Acc.**
SexWEs	**.406**	**.586**	.608	**.738**	**.786**
w/o phrase	.404	.571	**.611**	.726	.778
w/o refinement	.390	.536	.591	.713	.768
w/o RETROGAN	.398	.529	.594	.704	.760

Table 6: Results for SexWEs and ablative methods.

Performance versus Complexity Trade-off Analysis

According to experimental results, the overall performance of SexWEs fine-tuned by our cross-lingual domain-aware specialisation system shows 0.004-0.065 correlation score improvement in word similarity benchmark and 0.014-0.135 F1 score improvement in sexism detection. The results of both intrinsic and extrinsic evaluations demonstrate the effectiveness of specialised word vectors compared to pre-trained word vector baselines, and show improved performance over all other specialisation systems with similar model complexity. Compared with BERT-related baselines, our SexWEs is based on a simple TCNN architecture and still achieves a slight increase in the performance of detecting sexist content, showing further potential for more advanced and robust networks. Furthermore, we only need to train once to construct sexist word embeddings. Instead of only using it for sexism detection, it can also be reused to study sexism-related issues. Only by collecting new constraints, the methodology of building cross-lingual specialisation system can be further transferred to other low-resourced domains to detect abnormal behaviours online.

Conclusion and Future Work

To tackle sexism detection for low-resource languages, we propose an effective system for cross-lingual domain-aware semantic specialisation by injecting external constraints referring to sexist terms in both source and target languages. We report notable performance of SexWEs in both intrinsic and extrinsic evaluations, visualising the positive trend of word embeddings during the specialisation, as well as through an ablation study. However, we only observe a modest improvement after adding cross-lingual constraints, potentially due to its limited size. In the future, we plan to explore full automation of cross-lingual constraint creation and the extension of our approach to contextualised embeddings.

Ethical Considerations

Online sexism and abuse are sensitive subjects with various ethical concerns in the controversy surrounding the freedom of speech. To develop the fairness and reliability of our work, we address the following limitations:

- **Confidentiality:** Accessing the data is essential to make our work effective. Since the data is already public, to address the trade-off between privacy and effectiveness, original data has all personally identifiable data removed to ensure user anonymity.

- **Potential for harm:** Our work is not intended to harm vulnerable groups who are already discriminated against. While one could make bad use of sexism detection systems, such as learning to circumvent detection of their posts, our work is solely intended for the benign purposes of detection and mitigation of sexist speech.
- **Results communication:** Our work is free of plagiarism or research misconduct, but acknowledge potential limitations when analysing social media data, especially sexist data that does not clearly represent the attack target.

Acknowledgements

We would like to thank the anonymous reviewers for all their valuable comments and suggestions. Aiqi Jiang is funded by China Scholarship Council (CSC Funding, No. 201908510140). The authors acknowledge the support of the CONICET–Royal Society International Exchange (IEC\R2\192019). This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT https://doi.org/10.5281/zenodo.438045.

References

Alatawi, H. S.; Alhothali, A. M.; and Moria, K. M. 2021. Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT. *IEEE Access*, 9: 106363–106374.

Anzovino, M.; Fersini, E.; and Rosso, P. 2018. Automatic identification and classification of misogynistic language on twitter. In *Proc. of NLDB*, 57–64. Springer.

Arango, A.; Pérez, J.; and Poblete, B. 2021. Cross-lingual hate speech detection based on multilingual domain-specific word embeddings. *arXiv preprint arXiv:2104.14728*.

Arango, A.; Pérez, J.; and Poblete, B. 2020. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 101584.

Badjatiya, P.; Gupta, S.; Gupta, M.; and Varma, V. 2017. Deep learning for hate speech detection in tweets. In *Proc. of WWW*, 759–760.

Bassignana, E.; Basile, V.; and Patti, V. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proc. of CLiC-it*, volume 2253, 1–6. CEUR-WS.

Bigoulaeva, I.; Hangya, V.; Gurevych, I.; and Fraser, A. 2022. Addressing the Challenges of Cross-Lingual Hate Speech Detection. *arXiv preprint arXiv:2201.05922*.

Chiril, P.; Moriceau, V.; Benamara, F.; Mari, A.; Origgi, G.; and Coulomb-Gully, M. 2020. An Annotated Corpus for Sexism Detection in French Tweets. In *Proc. of LREC*, 1397–1403. ISBN 979-10-95546-34-4.

Colon-Hernandez, P.; Xin, Y.; Lieberman, H.; Havasi, C.; Breazeal, C.; and Chin, P. 2021. RetroGAN: A Cyclic Post-Specialization System for Improving Out-of-Knowledge and Rare Word Representations. In *Findings of ACL-IJCNLP*, 2086–2095.

Cui, Y.; Che, W.; Liu, T.; Qin, B.; Wang, S.; and Hu, G. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Findings of EMNLP*, 657–668.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*, 4171–4186.

Duchi, J.; Hazan, E.; and Singer, Y. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. 12(null): 2121–2159.

Faruqui, M.; Dodge, J.; Jauhar, S. K.; Dyer, C.; Hovy, E.; and Smith, N. A. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *Proc. of NAACL*, 1606–1615.

Fersini, E.; Nozza, D.; and Rosso, P. 2020. Ami@ evalita2020: Automatic misogyny identification. In *Proc. of EVALITA*.

Ganitkevitch, J.; Van Durme, B.; and Callison-Burch, C. 2013. PPDB: The Paraphrase Database. In *Proc. of NAACL*, 758–764.

Glavaš, G.; Litschko, R.; Ruder, S.; and Vulić, I. 2019. How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. In *Proc. of ACL*, 710–721.

Glavaš, G.; and Vulić, I. 2018a. Discriminating between Lexico-Semantic Relations with the Specialization Tensor Model. In *Proc. of NAACL*, 181–187.

Glavaš, G.; and Vulić, I. 2018b. Explicit Retrofitting of Distributional Word Vectors. In *Proc. of ACL*, 34–45.

Glick, P.; and Fiske, S. T. 2001. Ambivalent sexism. In *Advances in experimental social psychology*, volume 33, 115–188.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; and Mikolov, T. 2018. Learning Word Vectors for 157 Languages. In *Proc. of LREC*.

Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2016. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proc. of EMNLP*, 2116–2121. Austin, Texas: ACL.

Hewitt, S.; Tiropanis, T.; and Bokhove, C. 2016. The problem of identifying misogynist language on Twitter (and other online social spaces). In *Proc. of ACM WebSci*, 333–335.

Hill, F.; Reichart, R.; and Korhonen, A. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4): 665–695.

Jha, A.; and Mamidi, R. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proc. of NLP+CSS*, 7–16.

Jiang, A.; Yang, X.; Liu, Y.; and Zubiaga, A. 2022. SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27: 100182.

Jiang, A.; and Zubiaga, A. 2021. QMUL-SDS at EXIST: Leveraging Pre-trained Semantics and Lexical Features for Multilingual Sexism Detection in Social Networks. In *Proc. of IberLEF*, volume 2943, 469–483.

Jin, P.; and Wu, Y. 2012. SemEval-2012 Task 4: Evaluating Chinese Word Similarity. In *Proc. of *SEM*, 374–377.

Joulin, A.; Bojanowski, P.; Mikolov, T.; Jégou, H.; and Grave, E. 2018. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *Proc. of EMNLP*, 2979–2984.

Kamble, S.; and Joshi, A. 2018. Hate speech detection from code-mixed hindi-english tweets using deep learning models. *arXiv preprint arXiv:1811.05145*.

Kim, J.-K.; de Marneffe, M.-C.; and Fosler-Lussier, E. 2016. Adjusting Word Embeddings with Semantic Intensity Orders. In *Proc. of Repl4NLP*, 62–69.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proc. of EMNLP*, 1746–1751.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *Proc. of ICLR*.

Koufakou, A.; Pamungkas, E. W.; Basile, V.; and Patti, V. 2020. HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language. In *Proc. of the WOAH workshop*, 34–43.

Manne, K. 2017. *Down girl: The logic of misogyny*. Oxford University Press.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

Mrkšić, N.; Vulić, I.; Ó Séaghdha, D.; Leviant, I.; Reichart, R.; Gašić, M.; Korhonen, A.; and Young, S. 2017. Semantic Specialization of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints. *Transactions of the ACL*, 5: 309–324.

Navigli, R.; and Ponzetto, S. P. 2010. BabelNet: Building a very large multilingual semantic network. In *Proc. of ACL*, 216–225.

Pamungkas, E. W.; Basile, V.; and Patti, V. 2020. Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study. *Information Processing & Management*, 57(6): 102360.

Parikh, P.; Abburi, H.; Badjatiya, P.; Krishnan, R.; Chhaya, N.; Gupta, M.; and Varma, V. 2019. Multi-label Categorization of Accounts of Sexism using a Neural Framework. In *Proc. of EMNLP-IJCNLP*, 1642–1652.

Ponti, E. M.; Vulić, I.; Glavaš, G.; Mrkšić, N.; and Korhonen, A. 2018. Adversarial Propagation and Zero-Shot Cross-Lingual Transfer of Word Vector Specialization. In *Proc. of EMNLP*, 282–293.

Ponti, E. M.; Vulić, I.; Glavaš, G.; Reichart, R.; and Korhonen, A. 2019. Cross-lingual Semantic Specialization via Lexical Relation Induction. In *Proc. of EMNLP-IJCNLP*, 2206–2217.

Richardson-Self, L. 2018. Woman-Hating: On misogyny, sexism, and hate speech. *Hypatia*, 33(2): 256–272.

Rizos, G.; Hemker, K.; and Schuller, B. 2019. Augment to Prevent: Short-Text Data Augmentation in Deep Learning for Hate-Speech Classification. In *Proc. of ACM CIKM*, 991–1000. ACM New York, NY, USA.

Rodríguez-Sánchez, F.; Carrillo-de Albornoz, J.; and Plaza, L. 2020. Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. *IEEE Access*, 8: 219563–219576.

Rodríguez-Sánchez, F.; Carrillo-de Albornoz, J.; Plaza, L.; Gonzalo, J.; Rosso, P.; Comet, M.; and Donoso, T. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67: 195–207.

Samory, M.; Sen, I.; Kohne, J.; Flöck, F.; and Wagner, C. 2021. "Call me sexist, but..." : Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. In *Proc. of ICWSM*, volume 15, 573–584.

Shi, X.; and Zheng, Y. 2020. Perception and Tolerance of Sexual Harassment: An Examination of Feminist Identity, Sexism, and Gender Roles in a Sample of Chinese Working Women. *Psychology of Women Quarterly*, 44(2): 217–233.

Su, T.-r.; and Lee, H.-y. 2017. Learning Chinese Word Representations From Glyphs Of Characters. In *Proc. of EMNLP*, 264–273.

Sun, C.; Qiu, X.; and Huang, X. 2019. VCWE: Visual Character-Enhanced Word Embeddings. In *Proc. of NAACL*, 2710–2719.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Vulić, I.; Glavaš, G.; Mrkšić, N.; and Korhonen, A. 2018. Post-Specialisation: Retrofitting Vectors of Words Unseen in Lexical Resources. In *Proc. of NAACL*, 516–527.

Vulić, I.; Mrkšić, N.; and Korhonen, A. 2017. Cross-Lingual Induction and Transfer of Verb Classes Based on Word Vector Space Specialisation. In *Proc. of EMNLP*, 2546–2558.

Vulić, I.; Ponti, E. M.; Korhonen, A.; and Glavaš, G. 2021. LexFit: Lexical Fine-Tuning of Pretrained Language Models. In *Proc. of ACL-IJCNLP*, 5269–5283.

Wang, H.; Henderson, J.; and Merlo, P. 2021. Multi-Adversarial Learning for Cross-Lingual Word Embeddings. In *Proc. of NAACL*, 463–472.

Wang, X.; Jia, Y.; Zhou, B.; Ding, Z.-Y.; and Liang, Z. 2011. Computing semantic relatedness using chinese wikipedia links and taxonomy. *Journal of Chinese Computer Systems*, 32(11): 2237–2242.

Waseem, Z.; and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proc. of the NAACL SRW*, 88–93.

Wiegand, M.; Ruppenhofer, J.; Schmidt, A.; and Greenberg, C. 2018. Inducing a Lexicon of Abusive Words – a Feature-Based Approach. In *Proc. of NAACL*, 1046–1056.

Xiong, Z.; Qin, K.; Yang, H.; and Luo, G. 2021. Learning Chinese word representation better by cascade morphological n-gram. *Neural Computing and Applications*, 33(8): 3757–3768.

Zhang, T.; and You, F. 2021. Research on short text classification based on textcnn. In *Journal of Physics: Conference Series*, volume 1757, 012092. IOP Publishing.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proc. of ICCV*, 2223–2232.