Domain Adaptation for Text Classification with Weird Embeddings

Valerio Basile
University of Turin
valerio.basile@unito.it

Abstract

Pre-trained word embeddings are often used to initialize deep learning models for text classification, as a way to inject precomputed lexical knowledge and boost the learning process. However, such embeddings are usually trained on generic corpora, while text classification tasks are often domain-specific. We propose a fully automated method to adapt pre-trained word embeddings to any given classification task, that needs no additional resource other than the original training set. The method is based on the concept of word *weirdness*, extended to score the words in the training set according to how characteristic they are with respect to the labels of a text classification dataset. The *polarized weirdness* scores are then used to update the word embeddings to reflect task-specific semantic shifts. Our experiments show that this method is beneficial to the performance of several text classification tasks in different languages.

1 Introduction

In recent years, the Natural Language Processing community has directed a great deal of effort towards text classification, in different declinations. The list of shared tasks proposed at the recent editions (2016–2019) of the International Workshop on Semantic Evaluation (SemEval) shows an increasing number of tasks that can be cast as text classification problems: *given a text and a set of labels, choose the correct label to associate with the text*. If the cardinality of the set of labels is two, we speak of *binary* classification, as opposed to *multiclass* classification. Furthermore, not all binary classification tasks are the same. When the labels indicate the presence or absence of a given phenomenon, we speak of a *detection* task.

Classification tasks are mainly approached in a supervised fashion, where a labeled dataset is employed to train a classifier to map certain features of the input text to the probability of a certain label. Arguably, the most useful features in a NLP problem are the words that compose the text. However, in order to be processed by a machine learning algorithm, words need to be represented in a dense and machine readable format. *Word embeddings* solve this issue by providing vectorial representations of words where vectors that are close in the geometric space represent words that occur often in the same contexts. Among their applications, pre-trained word embeddings are a powerful source of knowledge to boost the performance of supervised models that aim at learning from textual instances.

Several deep learning models compute word embeddings at training time. However, they can be initialized with *pre-trained* word embeddings, typically computed on the basis of concordances in large corpora. This kind of initialization not only boosts the training of the model, but it also represents a way of injecting precomputed world knowledge into a model otherwise trained on a (sometimes very specific) data set.

An issue with word embedding models, including recent contextual embeddings such as Peters et al. (2018), is that they are typically trained on general-purpose corpora. Therefore, they may fail to capture semantic shifts that occur in specific domains. For instance, in a dataset of online hate speech, negatively charged words such as insults often co-occur with words that would normally be considered neutral, but carry instead a negative signal in that particular context. More concretely, in a dataset of hate speech towards immigrant in the post-Trump U.S., a word that otherwise would be considered neutral such as *wall* carries a definite negative connotation.

In this work, we try to capture this intuition computationally, and model this phenomenon in a word embedding space. We employ an automatic measure to score words in a labeled corpus according to their association with a given label (Section 3.1) and use this score in a fully automated method to adapt generic pre-trained word embeddings (Section 3.2). We test our method on existing benchmarks of hate speech detection (Section 4.1) and gender prediction (Section 4.2), reporting improvements in precision and recall.

2 Related Work

Kameswara Sarma et al. (2018) propose a method to adapt *generic* word embeddings by computing *domain specific* word embeddings on a corpus of text from the target domain and aligning the two vector spaces, obtaining a performance boost on sentiment classification. Another recent approach is based on projecting the vector representations from two domain-specific spaces into a joint word embedding model (Barnes et al., 2018b), building on a similar method applied to cross-lingual word embedding projection (Barnes et al., 2018a). With respect to these works, the approach proposed in this paper is significantly more lightweight, acting directly on a generic word embedding model without the need to train a domain specific one.

The word-level measure introduced in the next section is reminiscent of similar metrics from Information Theory, e.g., Information Content (Pedersen, 2010), and measures of frequency distribution similarity such as Kullback-Leibler divergence (Kullback and Leibler, 1951). However, in this paper we aimed at keeping the complexity of such computation low, in order to manually explore its effect on the word embeddings.

In the domain of hate speech, several approaches mix word embeddings and supervised learning with domain-specific lexicons (e.g., dictionaries of hateful terms), as highlighted by the description of participant systems to recent evaluation campaigns (Fersini et al., 2018; Bosco et al., 2018). These methods are computationally inexpensive, but require curated resources that are not always available for less represented languages.

3 Weirdness-based Embedding Adaptation

In this section, we present our method for automatic domain adaptation of pre-trained word embeddings. The input of the procedure is a set of pre-trained word embeddings and a corpus of texts paired with labels.

3.1 Polarized Weirdness

The Weirdness index was introduced by Ahmad et al. (1999) as an automatic metric to retrieve words characteristic of a *special language* with respect to their typical usage. According to this metric, a word is highly *weird* in a specific collection of documents if it occurs significantly more often in that context than in a general corpus. In practice, given a *specialist* text corpus and a *general* text corpus, the weirdness index of a word is the ratio of its relative frequencies in the respective corpora. Calling w_s the frequency of the word w in the specialist language corpus, w_g the frequency of the word w in the general language corpus, and t_s and t_g the total count of words the specialist and general language corpora respectively, the weirdness index of w is computed as:

$$Weirdness(w) = \frac{w_s/t_s}{w_g/t_g}$$

The weirdness index is used to retrieve words that are highly typical of a particular domain. For instance, in Ahmad et al. (1999), the words *dollar*, *government* and *market* are extracted from the TREC-8 corpus, a collection of governmental and financial domain, by comparing their frequencies to the general domain British National Corpus.

In this work, we propose a new application of the weirdness index to the task of text classification. Rather than comparing the frequencies of words from corpora of different domains, we compute the weirdness index based on the frequency of words occurring in labeled datasets. The mechanism is straightforward: instead of comparing the relative frequencies of a word in a special language corpus against a general language corpus, we compare the relative frequencies of a word as it occurs in the subset of a labeled dataset identified by one value of the label against its complement. Consider a labeled corpus $C = \{(e_1, l_1), (e_2, l_2), ...\}$ where $e_i = \{w_1, w_2, ...\}$ is an instance of text (e.g., an online comment), and

l_i is the label associated with e_i, belonging to a fixed set L (e.g., $\{positive, negative\}$).

The *polarized weirdness* (Florio et al., 2020) of w with respect to a specific label $l* \in L$ is the ratio of the relative frequency of w in the subset $\{e_i \in C : l_i = l*\}$ over the relative frequency of w in the subset $\{e_i \in C : l_i \neq l*\}$

Here is an example of how polarized weirdness is computed. Consider a corpus of 100 instances, 50 of which labeled *positive* and 50 labeled *negative*. The total number of words in instances labeled *positive* is 3,000, while the total number of words in instances labeled *negative* is 2,000. The word *good* occurs 50 times in *positive* instances and 5 times in *negative* instances. Therefore its polarized weirdness with respect to the positive label is:

$$PW_{positive}(good) = \frac{50/3,000}{5/2,000} = 6.66$$

However, the polarized weirdness of *good* with respect to the negative label is:

$$PW_{negative}(good) = \frac{5/2,000}{50/3,000} = 0.15$$

indicating that good is much more indicative of *positive*ness than *negative*ness.

Polarized weirdness can be computed at a low computational cost on any dataset labeled with categorical values, with just tokenization for preprocessing. The outcome of the calculation of the polarized weirdness index is a set of rankings, one for each label, over the vocabulary, there the top words in the ranking relative to a given label l are the most characteristic for that label.

3.2 Word Embedding Adaptation

In Section 3.1, we introduced an automatic metric that allows us to compute how much a word is characteristic to a certain label. We use this information to transpose the vector representing words highly typical of a label closer to each other in the vector space. Formally, once a label has been decided and the polarized weirdness is computed with respect to it, *for each pair of vectors* $\vec{v}_1, \vec{v}_2$ in a word embedding model, representing words with polarized weirdness pw_1 and pw_2 respectively, we compute new representations:

$$\vec{v}_1 = ((1 - \alpha \cdot pw_1)\vec{v}_1) + ((\alpha pw_2)\vec{v}_2)$$

$$\vec{v}_2 = ((1 - \alpha \cdot pw_2)\vec{v}_2) + ((\alpha pw_1)\vec{v}_1)$$

where α is a parameter controlling the extent of the adaptation. The result of the application of this algorithm is a new word embedding model over the same vocabulary as the original model, where pairs of word vectors are closer in the space to an extent proportional to their respective polarized weirdness score.

4 Experimental Evaluation

We test the word embedding adaptation introduced in Section 3 by adapting pre-trained multilingual word embeddings to three different tasks. For each task, the polarized weirdness index is computed on the labeled training sets as described in Section 3.1, and the generic word embeddings are adapted to the particular task domain applying the algorithm described in Section 3.2.

Our baseline model is a convolutional neural network (CNN) with a 64x8 hidden layer and Rectified Linear Units activation (ReLU), followed by a 4-size max pooling layer. We use the implementation from the Keras Python library[1], with ADAM optimization (Kingma and Ba, 2014), leaving the hyperparameters at their default value, except for optimization of learning rate (set between 10^{-2} and 10^{-3} depending on the dataset) and number of epochs (between 10 and 25).

We use the multilingual word embeddings provided by Polyglot (Al-Rfou et al., 2013). These are distributed word representations for over 100 languages trained on Wikipedia. The vector representations of words in Polyglot are 64-dimensional. The choice of this model is motivated by the need to have word embedding models for different languages that were created with the same method, to be able to measure improvements introduced merely by our adaptation method. In these experiments, we set $\alpha = 0.5$.

4.1 Experiment 1: Multilingual Hate Speech Detection

In the first experiment, the generic word embeddings are adapted to provide a better representation for words used in online messages containing hate speech towards women and immigrants. We use the dataset provided by the SemEval Task 5 (HatEval: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter), a public challenge where participants are invited to submit the predictions of systems for hate speech

[1] `https://keras.io/`

detection (Basile et al., 2019). In particular, we employ the data of the subtask A, where the prediction is binary (*hateful* vs. *not hateful*). The shared task website[2] provides datasets in Spanish and English, already divided into training, development and test sets. The topics of the messages are mainly two, namely women and immigrants, in a fairly balanced proportion. In fact, the dataset has been created by querying the Twitter API with a set of keywords crafted to capture these two topics. The English dataset comprises 13,000 tweets (10,000 for training and 3,000 for testing), with about 42% of the messages labeled as hateful. The Spanish dataset is smaller (6,600 tweets in total, 5,000 for training and 1,600 for testing), and it follows a similar distribution of topics and labels as the English set. Following are two examples of tweets from the English HatEval data,, with their Hate Speech label:

> I'd say electrify the water but that would kill wildlife. #SendThemBack
> `label: yes`
>
> Polish Prime Minister Mateusz Morawiecki insisted that Poland would push against any discussion on refugee relocations as part of the EU's migration politics.
> `label: no`

Similarly, two examples of tweets from the Spanish HatEval data, with translation and label:

> @rubenssambueza eres una basura de persona, lo cual no me sorprende porque eres SUDACA, y asi son los tercermundistas
> *@rubenssambueza you are garbage, which does not surprise me because you are a SUDACA, and so are third-worlders*
> `label: yes`
>
> Yo creía que ese jueguito solo existía para los árabes, jajaja.
> *I thought that this little game was only for arabs, ahahah.*
> `label: no`

The polarized weirdness of the words in the HatEval datasets (English and Spanish) is computed on the respective training sets as the ratio of their relative frequency in hateful messages over their relative frequency in non hateful messages. A modified version of the Polyglot embeddings is then computed[3] and the performance of the CNN using the adapted embeddings for initialization is compared with the performance obtained by initializing the CNN with the generic embeddings.

Table 1: Results of the English and Spanish Hate Speech Detection, for the negative (*no-HS*) and positive class (*HS*) and their macro-F1.

		no-HS			HS			Avg.
Model	**Acc.**	**Pr.**	**R.**	**F1**	**Pr.**	**R.**	**F1**	**F1**
English								
CNN	.468	.567	.401	.470	.398	.564	.466	.468
CNN+W	.482	.588	.394	.472	.413	.608	.492	.482
Spanish								
CNN	.528	.592	.595	.594	.437	.434	.436	.515
CNN+W	.527	.614	.497	.549	.450	.568	.502	.527

The results on the English dataset, presented in Table 1, show a clear improvement in the detection of hateful messages, leading to a +1.2% performance gain in macro-average F1-score. Recall is particularly impacted by the adapted embeddings, indicating that the modified model successfully helps in correcting false negatives.

The results on the Spanish HatEval task dataset, presented in Table 1 are even better than on English, with improvements in precision and recall for both the positive and the negative class, and a total gain of almost 2% macro-averaged F1-score. Similarly to English, the largest improvement is measured on the recall.

One of the advantages of the proposed method is that it is transparent with respect to the semantic shift computed on the pre-trained embeddings. Firstly, the words with the highest polarized weirdness index can be extracted, to gain insights into the specificity of the datasets. The top twenty weird words in the hateful English HatEval set are the following: nodaca, enddaca, kag, womensuck, @hillaryclinton, americafirst, trump2020, taxpayers, buildthewallnow, illegals, @senatemajldr, dreamer, buildthewall, they, @potus, walkawayfromdemocrat, votedemsout, wethepeople, illegalalien, backtheblue. The top twenty weird words in the hateful Spanish HatEval set with English translations are the following: mantero (*street vendor*), turista (*tourist*), negratas (*nigger*), caloría (*calory*), sanidad (*healthcare*), drogar (*to drug*), paises (*countries*), emigrante (*immigrant*), Hija (*daughter*), ZORRA (*bitch*), impuesto (*tax*), zorro (*bitch (masculine)*),

[2] https://competitions.codalab.org/competitions/19935

[3] To speed up to computation without major loss of information, we consider only the top 2,000 items from the weirdness ranking.

Table 2: Examples of words from the HatEval datasets, showing how their vector representation moves to reflect the semantic shift. Particular words that are generally neutral get closer to offensive words in the hate speech context.

Word embeddings	Generic word	Offensive word	Semantic shift	Cosine distance
Polyglot EN	wall	fuck	yes	1.224
Polyglot EN + P.W.	wall	fuck	yes	0.444
Polyglot EN	car	fuck	no	1.279
Polyglot EN + P.W.	car	fuck	no	1.413
Polyglot ES	directora (*director (F)*)	puta (*whore*)	yes	1.271
Polyglot ES + P.W.	directora (*director (F)*)	puta (*whore*)	yes	1.222
Polyglot ES	director (*director (M)*)	puta (*whore*)	no	1.366
Polyglot ES + P.W.	director (*director (M)*)	puta (*whore*)	no	1.411

totalmente (*totally*), lleno (*full*), invasor (*invader*), costumbre (*custom*), barrio (*neighborhood*), PAIS (*country*), Oye (*hey*), Españoles (*Spaniards*).

Secondly, one can extract the word embeddings after the polarized weirdness adaptation is applied, and qualitatively inspect their respective position in the vector space. Table 2 shows how certain pairs of words become more related in the adapted space, while others are untouched by the process. The example in Spanish is particularly interesting (and worrying), where a misogynistic derogatory word (*puta*) becomes closer to the feminine inflection of "director" but not to the masculine inflection.

4.2 Experiment 2: Gender Prediction

In the second experiment, we test our word embedding adaptation method in a different scenario, that is, the prediction of the gender of the author of messages. The assumption is that the most typical words used by each gender will cluster in the vector representation, thus helping the model discriminate them better.

We use the dataset distributed for the Cross-Genre Gender Prediction in Italian (GxG) shared task of the 2018 edition of EVALITA, the evaluation campaign of language technologies for Italian (Dell'Orletta and Nissim, 2018). The participants to the shared task are invited to submit the prediction of their system on a set of short and medium-length texts in Italian from different sources, including social media, news articles and personal diaries, on the gender of the author. The task is therefore a binary classification, evaluated by means of accuracy. We downloaded the data from the task website[4], comprising 22,874 instances divided into training set (11,000) and test set (10,874). The labels of the GxG are perfectly balanced between M (male) and F (female).

Table 3: Results of the Gender Prediction.

		Female			Male			Avg.
Model	**Acc.**	**Pr.**	**R.**	**F1**	**Pr.**	**R.**	**F1**	**F1**
CNN	.511	.507	.879	.643	.543	.143	.227	.435
CNN+W	.513	.508	.851	.636	.539	.174	.263	.450

Following are two examples of instances from the GxG dataset with their label and translation:

> @ElfoBruno no la barba la devo tenere lunga per sembrare folta perchè in realtà è rada...
> *@ElfoBruno no I have to keep the beard long to make it look thick because it really is patchy...*
> `label: M`

> Sabato prossimo sono davvero curiosa di scoprire cosa farà @Valerio_Scanu a #BallandoConLeStelle
> *Next Saturday I am very curious to find out what @Valerio_Scanu will do at #DancingWithTheStars*
> `label: F`

Since this is a *classification* rather than a *detection* task, the process is slightly different from the previous experiment, to account for the symmetry between the labels. First, the polarized weirdness is computed on the training set twice, once on the texts written by males (against the women's texts) and once on the texts written by females (against the men's texts). Then the general Polyglot embeddings are adapted by applying the algorithm in Section 3.2 twice, in both directions, using the respective weirdness rankings. The adapted embeddings are used to initialize the CNN, resulting

[4] https://sites.google.com/view/gxg2018/

in the classification performance presented in Table 3. The overall performance improves when the adapted embeddings are included in the model. However, the classification of the *male* label improves while the classification of *female* does not, due to the difference in recall.

Qualitative analysis reveals interesting patterns, confirming that strong bias is present in some pre-trained word embedding models. The twenty top weird words in the Male GxG set are: costituzionale (*constitutional*), socialisto (*socialist*), Lecce (*name of a city and a football club*), DALLA (*name of a singer*), utente (*user*), Samp (*name of a football team*), Sampdoria (*same of a football team*), Nera (*black*), allenatore (*coach*), Orlando (*proper name*), Bp (*acronym*), ni (*yes and no*), maresciallo (*marshall*), garanzia (*guarantee*), cerare (*to wax*), voluto (*willing*), pilotare (*to pilot*), disco (*disco*), caserma (*barracks*), From (*proper name*).

The top twenty weird words in the Female GxG set are instead the following: qualcuna (*someone (feminine)*), HEART EMOJI, Qualcuna (*someone (feminine)*), KISS EMOJI, 83 (*number*), essi (*them*), leonessa (*lioness*), Sarah (*proper name*), 06 (*number*), HEART-EYED EMOJI, nervoso (*nervous*), James (*proper name*), Dante (*proper name*), coreografia (*choreography*), Strada (*street*), Fra (*proper name*), Chiama (*call*), en (*en*), bravissimi (*very good (plural)*), Moratti (*proper name*). Arguably, a stronger topic bias (football) is present in the male subset, possibly explaining the better performance induced by the adaptation.

5 Conclusion and Future Work

In this work, we adapted an extension of the weirdness index to score the words in a labeled corpus according to how much they are typical of a given label. The polarized weirdness score is used to automatically adapt an existing word embedding space to better reflect target-specific semantic associations of words. We measured a performance boost on tasks of hate speech detection in English and Spanish, and gender prediction in Italian.

On detection tasks, the improvement from our method is remarkable in terms of recall, indicating the potential of weirdness-adapted word embeddings to correct false negatives. This result is in line with the original motivation for this approach, i.e., to account for semantic shift occurring in domain-specific corpora of opinionated content. For instance, in the hate speech domain, the adapted embeddings are able to capture that certain neutral words (e.g., “wall”) assume a polarized connotation (e.g., negatively charged).

The results from this study are promising, and encourage us to extend the method to richer representations (e.g., “weird” ngrams), languages other than European, and its integration into more sophisticated deep neural models. Recent Transformer models, in particular, compute contextualized embeddings, therefore including transformations similar to the present method. Although such models are less transparent with respect to such transformation, an experimental comparison is among the next steps planned in this research.

References

Khurshid Ahmad, Lee Gillam, and Lena Tostevin. 1999. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In *The Eighth Text REtrieval Conference (TREC-8)*, Gaithersburg, Maryland, November.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018a. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2483–2493. Association for Computational Linguistics.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018b. Projecting embeddings for domain adaption: Joint modeling of sentiment analysis in diverse domains. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 818–830. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*. Association for Computational Linguistics.

Cristina Bosco, Felice Dell’Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018.

Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018.*

Felice Dell'Orletta and Malvina Nissim. 2018. Overview of the EVALITA 2018 cross-genre gender prediction (gxg) task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018.*

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@SEPLN*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12).

Prathusha Kameswara Sarma, Yingyu Liang, and Bill Sethares. 2018. Domain adapted word embeddings for improved sentiment classification. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 51–59. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Ted Pedersen. 2010. Information content measures of semantic similarity perform better without sense-tagged text. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 329–332, Los Angeles, California, June. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.