A Computational Exploration of Pejorative Language in Social Media

Liviu P. Dinu[1], Ioan-Bogdan Iordache[1], Ana Sabina Uban[1], Marcos Zampieri[2]
[1]University of Bucharest, Romania
[2]Rochester Institute of Technology, USA
ldinu@fmi.unibuc.ro, iordache.bogdan1998@gmail.com
ana.uban@gmail.com, mazgla@rit.edu

Abstract

In this paper we study pejorative language, an under-explored topic in computational linguistics. Unlike existing models of offensive language and hate speech, pejorative language manifests itself primarily at the lexical level, and describes a word that is used with a negative connotation, making it different from offensive language or other more studied categories. Pejorativity is also context-dependent: the same word can be used with or without pejorative connotations, thus pejorativity detection is essentially a problem similar to word sense disambiguation. We leverage online dictionaries to build a multilingual lexicon of pejorative terms for English, Spanish, Italian, and Romanian. We additionally release a dataset of tweets annotated for pejorative use. Based on these resources, we present an analysis of the usage and occurrence of pejorative words in social media, and present an attempt to automatically disambiguate pejorative usage in our dataset.

1 Introduction

With the increase of social media usage, the issue of toxic language has become an important problem in our society. Automatic methods are needed to help mitigate this problem, and for this reason the study of toxic speech in NLP has become very popularity in recent years. Different categories and definitions have been proposed, including *hate speech* (Schmidt and Wiegand, 2017; Vashistha and Zubiaga, 2021), *offensive language* (Zampieri et al., 2019; Bucur et al., 2021), *aggression* (Kumar et al., 2018, 2020), as well as further sub-categories depending on the targets, such as women, migrants, etc. (Basile et al., 2019). From a computational perspective, the problem is usually approached as a classification task at the post level, where a classifier is trained to predict whether a social media post contains offensive/toxic language.

In this paper we address the question of **pejorative** words. Pejorative words are *words or phrases that have negative connotations or that are intended to disparage or belittle*[1]. Pejorativity is closely related to the notion of slurs or insults: "as noun phrases, 'insult' and 'slur' refer to symbolic vehicles designed by convention to derogate targeted individuals or groups" (Anderson and Lepore, 2013). While pejorative language is often used in offensive speech (Castroviejo et al., 2020), they are not identical categories. There are offensive posts that do not use pejorative words (e.g. "Women belong in the kitchen"), and pejorative uses of words that are not harmful ("What a *shitty* chair") because the offensive content is not targeted at a person or a group as described in the popular annotation taxonomy of the Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019).

Words can have a negative meaning in one context and not in others (such as the figurative meanings of "trash" or "pussy"); or be pejorative in one language or culture, and not in others (such as the Romanian "cioara" (literally, "crow") - a slur for people of color). Slurs can also lose their pejorative meaning through semantic change (e.g. the word "queer" went through semantic amelioration over the years - it used to be a slur and is losing its negative connotation (Brontsema, 2004)). Recognizing the complexity of the phenomenon, with its linguistic subtleties as well as the variability related to culture and context, are important to successfully recognize pejorative words and by extension offensive posts and hate speech.

Pejorative language is still largely under-explored in computational linguistics. There are very few studies addressing or taking pejorative language into account (Wiegand et al., 2018; Mendelsohn et al., 2020; Palmer et al., 2017; Eder et al., 2019; Castroviejo et al., 2020). A few related works

[1]https://www.merriam-webster.com/dictionary/pejorative

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3493–3498
November 7–11, 2021.

to ours include Palmer et al. (2017) who focused on pejorative connotations for nominalized adjectives and Mendelsohn et al. (2020) who built a lexicon of vulgar terms (and vulgarity scores) for German based on derogatory terms found in Wiktionary.

In this study, we address this important gap by leveraging dictionaries to build a multilingual lexicon of pejorative language for four languages. We compare the occurrence of pejorativity in social media with other established categories of toxic language, relying on existing hate speech corpora. Unlike most existing studies in hate speech and offensive language identification, our paper focuses on the lexical level and approaches the issue of ambiguity in toxic language, formulating the problem of pejorativity detection as a word sense disambiguation (WSD) task. The main contributions of this work are the following:

1. We create a multilingual lexicon of pejorative words in four languages: English, Spanish, Italian, and Romanian.
2. We present several experiments to automatically distinguish pejorative from non-pejorative uses of words relying on state-of-the-art word sense representations based on contextual embeddings.
3. We release annotated datasets containing pejorative words in English and Spanish tweets.

2 Pejorative Lexicon

2.1 Data Collection

We started by gathering a pejorative lexicon for four languages: English, Spanish, Italian and Romanian. For each language, we assembled a list of words that can be used with a pejorative sense according to existing language resources. We focused on providing a lexicon consisting of words that can be used pejoratively on their own, rather than words that are part of pejorative expressions or idioms. In order to collect these terms for English, Spanish, and Italian we used Wiktionary[2], and collected the terms that were part of the "derogatory terms" category. For Romanian, we used another online-available dictionary, *dexonline*[3], and selected all of the words that had a pejorative definition and where the definition was intended for the word not for an expression built around the word.

[2] https://www.wiktionary.org/
[3] https://dexonline.ro/

2.2 Lexicon Description

For each language's lexicon, we computed the frequency of each word, based on occurrence across different large corpora including Wikipedia and social media datasets, using the *wordfreq* Python library (Speer et al., 2018). We used the WordNet (Miller, 1995) to count the number of senses a word can have (by counting the number of *synsets* that they are contained in) as well as their parts of speech. Statistics are shown in Table 1. The distribution across parts of speech is illustrated in Figure 1. For a given word, we counted all its possible parts of speech according to WordNet.

Lang.	Words	WF cover.	WN cover.	Senses
EN	2903	28.97%	25.56%	3.07
ES	881	51.99%	18.05%	3.05
IT	149	53.02%	49.66%	1.87
RO	770	12.34%	32.21%	2.41

Table 1: Number of words for each language, coverage in *wordfreq*, WordNet coverage, and average number of senses for words in WordNet.

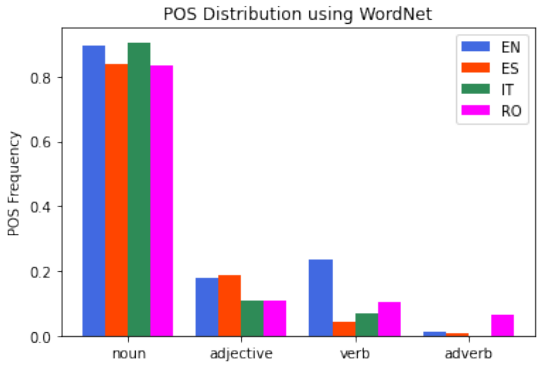

Figure 1: Distribution of parts of speech for the collected words for each language in WordNet.

3 Pejorative Tweet Dataset

For building a data set of English texts containing words that are used pejoratively, we started by looking at three datasets of hate speech on Twitter: (Davidson et al., 2017), (Basile et al., 2019). (Waseem and Hovy, 2016), and selected the tweets that contain words from our pejorative lexicon (after normalizing words to their stems). For each data set, we extracted pairs of words and tweets where they occur.

The dataset published by Davidson et al. (2017) contains tweets annotated with one of three classes (hateful, offensive and neither). For each label, the

number of pejorative words found in the tweets is the following: $1,114$ out of $1,430$ hateful tweets, $8,358$ out of $19,190$ offensive tweets, and $2,221$ among the remaining $4,163$ tweets were found to contain pejorative words. The hate speech dataset published as part of the HatEval shared task (Basile et al., 2019) contains tweets annotated with labels for hateful and aggressive speech. Out of the $4,210$ hateful tweets, $1,985$ contain words from our lexicon, while from $1,763$ aggressive tweets, 822 were selected. Finally, the dataset by Waseem and Hovy (2016) contains tweets annotated for racist and sexist speech. 8 tweets out of the $1,970$ racist tweets, and 897 from $3,378$ sexist tweets, contain pejorative words.

For Spanish, we employed the same technique of filtering tweets. We looked at the Spanish tweets data set provided by Basile et al. (2019) and considered only the binary label for hate speech classification. Out of the total of $5,000$ tweets, we have extracted $1,621$ hateful examples and $1,667$ non-hateful examples that contain words from our Spanish pejorative lexicon.

3.1 Annotation

We then built a data set of English tweets annotated for pejorative usage of words, by selecting tweets from the HatEval data set (Davidson et al., 2017), which we chose given the large number of unique pejorative words it contains ($1,77$ for hate, $3,95$ for offensiveness and $2,77$ for none). We extracted two separate data sets in two different ways.

The first data set (PEJOR1) was built by selecting a fixed percentage of tweets from each class, in order to obtain a balanced dataset with respect to the three labels (keeping only words that are represented at least once in each class). In this way, we attempt to conserve the relative distribution of the pejorative stems across the three classes.

The second data set (PEJOR2) was built to be balanced with regard to both the words' distribution and the original labels. For each pejorative stem we extracted a fixed number of pairs from each of the three classes.

The selected tweet-word pairs extracted for both of the data sets were then annotated with binary valued labels, denoting whether the word in the pair is used pejoratively (label 1) or not (label 0) in the tweet. We used the Wiktionary definitions in order to label words as pejorative only when used with senses marked as "derogatory" in Wiktionary.

The Table 2 shows statistics for the two datasets, while Figure 2 illustrates the distribution of labels for words in PEJOR2. Data was annotated by specialists in linguistics. We used two annotators for each datapoint, and used a third one where there was disagreement. The obtained Cohen's k agreement score was 0.933.

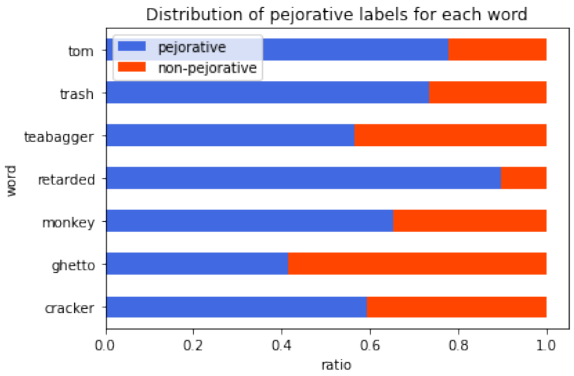

Figure 2: Distribution of labels for the PEJOR2 English dataset.

PEJOR1				PEJOR2			
	pairs	words	label 1		pairs	words	label 1
	944	23	49.7%		313	11	51.4%
	hate	offensive	neither		hate	offensive	neither
0	8.04%	21.59%	20.74%	0	12.46%	15.34%	20.77%
1	27.20%	14.07%	8.36%	1	21.09%	17.89%	12.46%

Table 2: Number of tweet-word pairs in the datasets, number of unique words, and the frequency of the 1 label. Overlap with (Davidson et al., 2017) labels.

For Spanish, we built a pejorative data set by selecting tweets from the (Basile et al., 2019) data set, following the same approach used for extracting the PEJOR2 English examples. We annotated a small subset of the tweets, consisting of 12 pejorative words with 10 tweets each (balanced between hateful and non-hateful tweets).

4 Classification Experiments

The classification task we approached was inferring the 0/1 label for tweet-word pairs. Namely, given a word and a tweet, where the word appears in the tweet, we want to be able to say if the word was used pejoratively or not in that tweet.

In order to prepare our data, for each tweet-word pair, the tweet was tokenized and the position of the occurrence of the word was found among the tokens. Then, we generated a contextual embedding (Devlin et al., 2019) for that occurrence, by employing various BERT models, pre-trained on

English texts, provided by the *huggingface* Python library (Wolf et al., 2019). The embedding obtained for the specified position is computed by summing the 768-dimensional hidden states generated for that position by each of the 12 layers of the BERT architecture. We note that, for out-of-vocabulary words, the BERT tokenizer provided by the *huggingface* library splits them into sub-words. In this case we chose to generate the embeddings for each of the sub-words of our word occurrence and then average them to obtain the final 768-dimensional embedding.

Figure 3 illustrates an example of uses of a pejorative word ("cracker") in the PEJOR2 dataset, by representing its embeddings reduced to two dimensions using PCA. We can see that most of the similar labelled examples are clustered together.

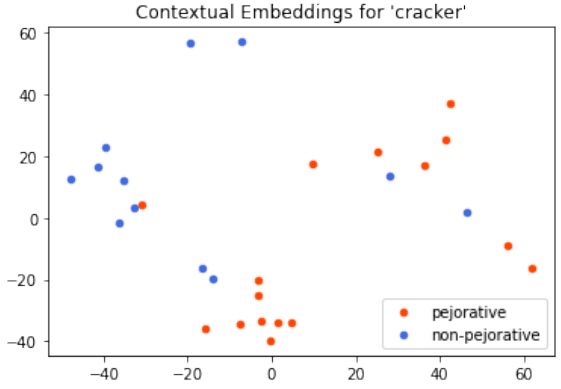

Figure 3: 2D plot of the contextual embeddings generated for the word 'cracker' in the PEJOR2 data set, for each of its occurrences in the tweets, using a pretrained BERT model. Embeddings were reduced to two dimensions using PCA.

For classification on our English data set, we grouped the pairs by the pejorative word contained in the tweet, and independently for each group, we fitted a classifier on the contextual embeddings (Liu et al., 2020). For extracting the embeddings we used various transformer models (BERT base (Devlin et al., 2019), BERTweet (Nguyen et al., 2020), RoBERTa (Liu et al., 2019), Multilingual BERT (Devlin et al., 2019)) and for the classification algorithm we used K-Nearest Neighbors, Support Vector Machines (SVM), Multilayer Perceptron (MLP). For K-Nearest Neighbors, we considered the cosine similarity as the distance function and found through hyper-parameter tuning that neighborhoods of size 4 were the best performing setting.

For evaluation, we employed a 5-fold cross-validation. Performance metrics were computed for each word independently, measuring the capacity of distinguishing the pejorative and non-pejorative usage of the word in different contexts. We report, for each metric, the value resulted by averaging over the scores obtained for all of the word groups. We leave out from this averaging the words that appear with only one label in the whole data set (only pejorative or only non-pejorative), since they will be always classified correctly regardless of the contextual embeddings. We also employed a baseline that based on the training data it learns to predict only the most frequent label. Table 3 shows the obtained results. The appendix contains a table with nearest neighbors found for example tweets.

We notice a promising performance of the classifiers in distinguishing pejorative usage, of up to 0.86 F1-score. Following the best performing models for each data set, overall 107 samples were misclassified in the PEJOR1 dataset, while for PEJOR2 there were 37. Words in PEJOR2 seem slightly easier to classify, which might be expected given the dataset is more balanced in positive and negative examples.

Dataset		PEJOR1		PEJOR2	
Embeddings	Classifier	Acc	F1	Acc	F1
—	baseline	67.7%	0.604	67.3%	0.694
BERT base	4-NN	76.9%	0.776	81.1%	0.841
BERT base	SVM	79.2%	0.768	80.3%	0.837
BERT base	MLP	79.8%	0.801	82.5%	**0.864**
RoBERTa	4-NN	72.6%	0.724	67.7%	0.716
RoBERTa	SVM	72.1%	0.654	68.9%	0.692
RoBERTa	MLP	76.4%	0.781	77.2%	0.802
BERTweet	4-NN	80.4%	0.797	75.4%	0.776
BERTweet	SVM	78.0%	0.760	77.9%	0.793
BERTweet	MLP	81.9%	**0.802**	78.1%	0.803
Multilg. BERT	4-NN	71.0%	0.714	74.2%	0.784
Multilg. BERT	SVM	73.0%	0.657	74.3%	0.786
Multilg. BERT	MLP	76.9%	0.750	75.1%	0.796

Table 3: Performance scores for various contextual embeddings and classifiers on the PEJOR1 and PEJOR2 English data sets

For the Spanish pejorative data set, since most of the examples were not labelled, we tried an unsupervised clustering approach. For each group of example pairs defined by the common pejorative word, we extracted contextual embeddings using the same previously explained method. Using KMeans clustering, we grouped those embeddings into two classes. We then computed, using the annotated examples, the amount of overlap between those two clusters and the pejorative la-

bels. The overlap was computed as the accuracy and the macro-F1 score of the clusters when used for predicting the labels. We averaged the scores computed for all of the groups where there was at least one positively and one negatively labelled example. The results obtained using various embeddings (BETO (Cañete et al., 2020) and Multilingual BERT (Devlin et al., 2019)) can be found in table 4. For reference, we have used the random chance of assigning the clusters as a baseline.

Method	Accuracy	F1 score
random chance	50.0%	0.488
BETO	**68.9%**	**0.573**
Multilingual BERT	65.0%	0.503

Table 4: Overlap score for unsupervised clustering on the Spanish pejorative data set

5 Conclusions

We have addressed an important but under-explored lexical category in the intersection of lexical semantics and toxic speech: pejorativity. We released a public lexicon of pejorative words in four languages (including a low-resource language), as well as dataset of tweets annotated for pejorative uses of words.[4] We have modelled pejorativity detection as a problem of disambiguation, and performed experiments using state-of-the-art contextual embeddings in order to automatically distinguish pejorative from non-pejorative uses of words, obtaining promising results. In the future, we would like to explore modelling the problem of pejorativity detection as a sequence labelling task.

At the application level, integrating pejorativity detection into hate speech detection systems, for example, would be a promising area for future research. From a linguistic perspective, it would be interesting to analyze occurrence and pejorative value cross-lingually taking advantage of large pretrained cross-lingual models as in Ranasinghe and Zampieri (2020, 2021) for offensive language identification. We expect pejorative connotations to be difficult to translate and not transfer well across languages, which could also have practical implications. We would also like to extend our dataset of social media posts to cover more pejorative terms, as well as other languages.

[4]The lexicon and the corpus are available at: https://nlp.unibuc.ro/resources

Ethical Considerations

Our dataset of tweets was obtained by sampling existing hate and offensive speech datasets cited in this paper, complying with the terms of use of each of these datasets. All datasets were anonymized, no usernames or any of their demographics are included in the data used to train our models.

Acknowledgments

We warmly thank our annotators Laurenţia Nodit and Laurenţiu Zoicaş for their time. We would like to thank the anonymous EMNLP reviewers for their insightful feedback.

This research is supported in part by the Romanian Ministry of Research, Innovation and Digitization, CNCS/CCCDI UEFISCDI, project number 411PED/2020 and project number 108PCE/2021, within PNCDI III.

References

Luvell Anderson and Ernie Lepore. 2013. What did you call me? slurs as prohibited words setting things up. *Analytic Philosophy*, 54(3):350–63.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of SemEval*.

Robin Brontsema. 2004. A queer revolution: Reconceptualizing the debate over linguistic reclamation. *Colorado Research in Linguistics*, 17.

Ana-Maria Bucur, Marcos Zampieri, and Liviu P. Dinu. 2021. An exploratory analysis of the relation between offensive language and mental health. In *Findings of the ACL*.

Elena Castroviejo, Katherine Fraser, and Agustín Vicente. 2020. More on pejorative language: Insults that go beyond their extension. *Synthese*, pages 1–26.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings ICWSM*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep

bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2019. At the lower end of Language—Exploring the vulgar and obscene side of German. In *Proceedings of the ALW*.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of TRAC*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of TRAC*.

Jerry Liu, Nathan O'Hara, Alexander Rubin, Rachel Draelos, and Cynthia Rudin. 2020. Metaphor detection using contextual word embeddings from transformers. In *Proceedings of Fig-Lang*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3:55.

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

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Alexis Palmer, Melissa Robinson, and Kristy K. Phillips. 2017. Illegal is not a noun: Linguistic form for detection of pejorative nominalizations. In *Proceedings of ALW*.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of EMNLP*.

Tharindu Ranasinghe and Marcos Zampieri. 2021. MUDES: Multilingual Detection of Offensive Spans. In *Proceedings of NAACL*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of SocialNLP*.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. Luminosoinsight/wordfreq: v2.2.

Neeraj Vashistha and Arkaitz Zubiaga. 2021. Online multilingual hate speech detection: experimenting with hindi and english social media. *Information*, 12(1):5.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of NAACL SRW*.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of NAACL*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL*.