HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language

Anna Koufakou♣ **Endang Wahyu Pamungkas**♡ **Valerio Basile**♡ **Viviana Patti**♡
♣Florida Gulf Coast University, Software Engineering Dept, USA
♡University of Turin, Dipartimento di Informatica, Italy
♣akoufakou@fgcu.edu ♡{pamungka,basile,patti}@di.unito.it

Abstract

The detection of abusive or offensive remarks in social texts has received significant attention in research. In several related shared tasks, BERT has been shown to be the state-of-the-art. In this paper, we propose to utilize lexical features derived from a hate lexicon towards improving the performance of BERT in such tasks. We explore different ways to utilize the lexical features in the form of lexicon-based encodings at the sentence level or embeddings at the word level. We provide an extensive dataset evaluation that addresses in-domain as well as cross-domain detection of abusive content to render a complete picture. Our results indicate that our proposed models combining BERT with lexical features help improve over a baseline BERT model in many of our in-domain and cross-domain experiments.

1 Introduction

The automatic classification of abusive and offensive language is a complex problem, that has raised a growing interest in the Natural Language Processing community in the last decade or so (Fortuna and Nunes, 2018; Vidgen et al., 2019; Poletto et al., 2020). Several benchmarks have been introduced to measure the performance of mostly supervised machine learning systems tackling such problems as text classification tasks (Basile et al., 2019; Zampieri et al., 2019b). The evaluation of abusive and offensive language, however, is not straightforward. Among the issues, it has been observed how the topics discussed in the messages composing the benchmark datasets introduce biases, interfering with the modeling of the pure pragmatic phenomena by the supervised models trained on the respective training sets (Wiegand et al., 2019; Caselli et al., 2020).

Among the recent neural architectures, BERT (*Bidirectional Encoder Representations from Transformers* (Devlin et al., 2019)), is considered the state of the art in several NLP tasks, including abusive and offensive language detection. For example, in the SemEval 2019 Task 6 (Zampieri et al., 2019b, OffensEval), seven out of the top-ten teams used BERT, including the top team. The knowledge encoded in such model, based on *transformer* neural networks, is induced by a pre-training performed on a large quantity of text, then fine-tuned to a specific dataset in order to learn complex correlations between the natural language and the labels. One disadvantage to models such as BERT is that no additional external knowledge is taken into consideration, such as linguistic information from a lexicon.

In this paper, we propose a hybrid methodology to infuse external knowledge into a supervised model for abusive language detection. We propose to add extra lexical features with BERT at sentence- or term-level, with the goal of improving the quality of its prediction of abusive language. In particular, we investigate the inclusion of features from a categorized lexicon in the domain of offensive and abusive language, with the aim of supporting transfer knowledge in that domain across datasets.

We perform extensive, in-domain and cross-domain experimentation, to evaluate the performance of models which are trained on one dataset and tested on other datasets. Cross-domain classification of abusive content has been proposed to address the diverse topical focuses and targets as exhibited in different datasets developed from the research community in the last years (Karan and Šnajder, 2018; Pamungkas and Patti, 2019; Pamungkas et al., 2020b). For example, some datasets proposed for hate speech detection focus on racism or sexism (Waseem and Hovy, 2016), in line with the target-oriented nature of hate speech, while others on offensive or abusive language without tar-

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 34–43
Online, November 20, 2020.
https://doi.org/10.18653/v1/P17

geting a specific vulnerable group (Zampieri et al., 2019a; Caselli et al., 2020). This makes it difficult to know if a model that performs well on one dataset will generalize well for other datasets. However, several actors – including institutions, NGO operators and ICT companies to comply to governments' demands for counteracting online abuse[1]– have an increasing need for automatic support to moderation (Shen and Rose, 2019; Chung et al., 2019) or for monitoring and mapping the dynamics and the diffusion of hate speech dynamics over a territory (Paschalides et al., 2020; Capozzi et al., 2019) considering different targets and vulnerable categories. In this scenario, there is a considerable urgency to investigate computational approaches for abusive language detection supporting the development of robust models, which can be used to detect abusive contents with different scope or topical focuses. When addressing this challenge, the motivation for our proposal is the hypothesis that the addition of lexical knowledge from an abusive lexicon will soften the topic bias issue (Wiegand et al., 2019), making the model more stable against cross-domain evaluation. Our extensive experimentation with many different datasets shows that our proposed methods improve over the BERT baseline in the majority of the in-domain and cross-domain experiments.

2 Related Work

The last ten years have seen a rapidly increasing amount of research work on the automatic detection of abusive and offensive language, as highlighted by the success of international evaluation campaigns such as HatEval (Basile et al., 2019) on gender- or ethnic-based hate speech, OffensEval (Zampieri et al., 2019b, 2020) on offensive language, or AMI (Fersini et al., 2018a,b, Automatic Misogyny Identification) on misogyny. Several annotated corpora have also been established as benchmarks besides the data produced for the aforementioned shared tasks for several languages, for instance, Waseem et al. (2017) for racism and sexism in English, Sanguinetti et al. (2018) for hate speech Italian and Mubarak et al. (2017) for abusive language in Arabic. We refer to (Poletto et al., 2020) for a systematic and updated review of resources and benchmark corpora for hate speech detection across different languages.

The vast majority of approaches proposed in the literature are based on supervised learning, with statistical models learning the features of the target language and their relationship with the abusive phenomena from an annotated corpus. Most works propose variations on neural architectures such as Recurrent Neural Networks (especially Long Short-term Memory networks), or Convolutional Neural Networks (Mishra et al., 2019). An investigation on what type of attention mechanism (contextual vs. self-attention) is better for abusive language detection using deep learning architectures is proposed in (Chakrabarty et al., 2019). Character-based models have also been proposed for this task (Mishra et al., 2018).

More recently, models based on the *Transformer* neural network architecture have gained prominence, thanks to their ability of learning accurate language models from very large corpora in an unsupervised fashion, and then being fine-tuned to specific classification tasks, such as abusive language detection, with relatively little amount of annotated data. Several ideas have been proposed in the literature to improve the performance of BERT for abusive language detection. For example, fine-tuning large pre-trained language models in (Bodapati et al., 2019).

A complementary approach to supervised learning towards the detection of abusive and offensive language is the use of language resources such as lexicons and dictionaries. Wiegand et al. (2018) proposed a method to induce a list of English words to capture abusive language. Davidson et al. (2017) introduced an English lexicon covering hate speech, racism, sexism, and homophobia. Other languages have relatively less resources with respect to English, apart perhaps from Arabic, for which two lexical resources are available, by Mubarak et al. (2017), with focus on obscenity, and by Albadi et al. (2018). A notable exception is HurtLex (Bassignana et al., 2018), a multilingual lexicon of offensive words, created by semi-automatically translating a handcrafted resource in Italian by linguist Tullio De Mauro (called *Parole per Ferire*, "words to hurt" (De Mauro, 2016)) into 53 languages. Lemmas in HurtLex are associated to 17 non-mutually exclusive categories, plus a binary macro-category indicating whether the lemma reflects a stereotype. The number of lemmas in any language of HurtLex is in the order of thousands, depending on the lan-

[1] See for instance the Code of Conduct on countering illegal hate speech online issued by EU commission (EU Commission, 2016).

guage, and they are divided into the four principal parts of speech: noun, adjective, verb, and adverb. In our earlier research, we used a technique called *retrofitting* to enhance word embeddings using HurtLex, for detecting aggression in English, Hindi, and Bengali (Koufakou et al., 2020).

In this work, we propose to infuse the lexical knowledge from HurtLex into a BERT model with the goal to improve the efficacy of abusive and offensive language prediction models. Specifically, we utilize different representations of the HurtLex categories as they are found in the data, utilize them with a BERT model, and explore how they affect the detection accuracy. To the best of our knowledge, the utilization of a hate lexicon, especially one that is based on this kind of structure with multiple categories, with a BERT model has not been explored before. We fully describe our methods in the following section.

3 Methodology

In this paper, we explore two models based on how they utilize the lexical features extracted from the hate speech lexicon, HurtLex (Bassignana et al., 2018). Both of our proposed models utilize two inputs: (a) the sentence tokens (BERT's usual input), and (b) a vector we create based on the categories in HurtLex as they are found in our data. All the data we explore in this work are in English, so we used only the English version of HurtLex and leave the multilingual aspect for future work. Specifically, we use the English section of HurtLex version 1.2 [2]. It contains 6,072 entries, of which 2,268 are in the *conservative* subset (these are terms with higher confidence). Table 1 lists the categories in HurtLex, with the number of terms in each one, as well as examples.

Our models both start using the BERT layer, which takes three inputs consisting of id, mask and segment - see Figures 1 and 2. The output of this BERT layer connects to a dense layer. Please note that specific details and parameters for the BERT Baseline as well as any layers in our models are presented in section 4.

Regarding HurtLex features, we have two ways of extracting features: encodings and embeddings. In the first architecture (see Figure 1), based on the words in the train set, we find their categories in HurtLex and then derive a vector of HurtLex categories: we call this HurtLex Encoding. The total number of categories in HurtLex is 17, so the dimensionality of the HurtLex encoding is 17. Each element in this vector is simply a frequency count for the respective category in HurtLex. For example, if there is a total of 3 words in a train record (e.g. tweet) that belong in the *ethnic slurs* category of HurtLex, then the corresponding element in the HurtLex encoding is 3. We call this architecture *HurtBERT-Enc*.

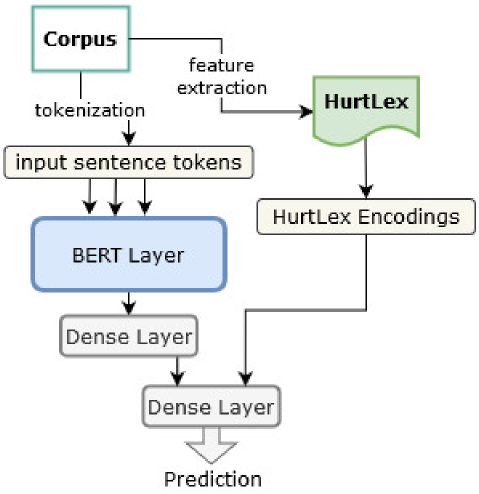

Figure 1: HurtBERT-Enc, our model using HurtLex Encodings

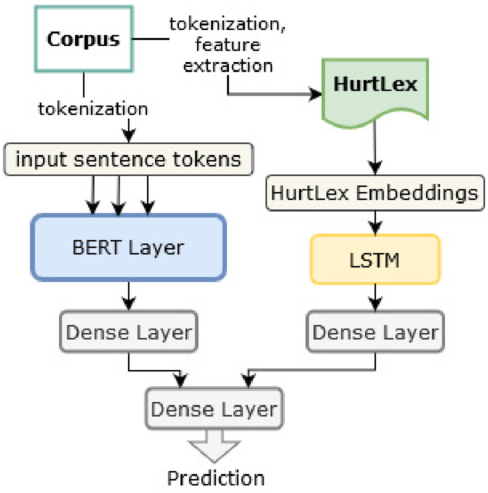

Figure 2: HurtBERT-Emb, our model using HurtLex Embeddings

Our second model explores using HurtLex embeddings with an LSTM, as shown in Figure 2. The HurtLex embedding is a 17-dimension one-hot encoding of the word presence in each of the lexicon categories. This model is named *HurtBERT-Emb*.

One of the main differences between the embedding and the encoding is that the embedding

[2] https://github.com/valeriobasile/hurtlex

Category	# Terms	Examples
negative stereotypes ethnic slurs	371	barbarian, idiotic, dummy, n***oes, infertility
locations and demonyms	24	genoan, savage, barbarian, tike, boor
professions and occupations	192	wooer, politician, peasant, fishwife, academism
physical disabilities and diversity	63	handycapped, midget, worthless, invalidity, impaired
cognitive disabilities and diversity	491	artless, retarded, simple, goof, brute
moral and behavioral defects	715	close-minded, cheater, stinking, forgery, faker
words related to social and economic disadvantage	124	miscreants, miserable, wretch, pitiful, villain
plants	177	finocchio, potato, papaya whip, squash, f**ot
animals	996	b***h, t**t, goose, scoundrel, beastly
male genitalia	426	wanky, c**k, testicles, phallic, prick
female genitalia	144	babe, c**t, t**t, boob, p***y
words related to prostitution	276	s*ut, street walker, crack h*, hooker, w***e
words related to homosexuality	361	drag, crossdressing, shirtlifter, f**, qu**rio
with potential negative connotations	518	bollocks, acolyth, delirious, reject, mooch
derogatory words	2,204	scalawag, boaster, rustler, dunderheaded, pedant
felonies and words related to crime and immoral behavior	619	mafioso, roguery, robber, scalawag, rapscallion
words related to the seven deadly sins of the Christian tradition	527	concupiscience, laziness, vanity, madness, slacker

Table 1: Descriptions, number of terms and examples for the categories in HurtLex

is word-level, while the encoding is comment-level. Therefore, for the case of HurtLex encodings, one record (e.g. one tweet) generates one 17-dimensional vector (which we call HurtLex encoding). While, for the case of the HurtLex embeddings, every word in the comment has one 17-dimensional vector representation (which is the HurtLex embedding). The HurtLex embedding also passes through an embedding layer, which goes into an LSTM and a dense layer, as shown in Figure 2.

In the end, the encoding is a simple representation that reflects if a category from the lexicon is found in the words of comment (or more accurately, how many times this category is found). While the embedding-based model also represents non-linear interactions between the features, that is, linguistically, the role of the HurtLex words in the sentence.

Finally, for both models, we concatenate the dense layer from the BERT output and the dense layer from the HurtLex output, before passing into a dense layer with sigmoid activation as the predictor layer (see the bottom part of both Figures 1 and 2).

4 Experiments and Results

Overall, we follow the experimental setup of (Swamy et al., 2019). We exploit the BERT pre-trained models available on tensorflow-hub[3], which facilitate us to integrate BERT on top of Keras architecture[4]. Specifically, we use the `bert-uncased` model, with 12 transformer blocks, 12 self-attention heads, and hidden layer dimension 768. Based on performance in early experiments, our models use learning rate of $e^{-}5$, batch size 32, and maximum sequence length of 50. We implement early stopping and model checkpoint based on the development set evaluation to avoid overfitting during the training process. For the LSTM in Figure 2, we use 32 nodes, and the dense layers in Figures 1-2 are 256 and 16 nodes respectively (all dense layers except last have RELU activation).

4.1 Datasets

The datasets used in our experiments are summarized in Table 2. All the datasets we explore in this work are in English: we leave the multilingual aspect of this research to future work. Similar to previous work in cross-domain classification of abusive language, all datasets need to be cast into binary label as abusive (in bold in the Table) and not abusive. We split all datasets into training, development and test sets with the proportion of 70%, 10% and 20% respectively. We list and describe the datasets below in chronological order, as some of the datasets were built based on previous data or annotation schemes.

Waseem: This corpus was collected over a period of 2 months by using representative keywords which is frequently used to attack specific targets including religious, sexual, gender and ethnic minorities (Waseem and Hovy, 2016). Two annotators were assigned to annotate the full dataset, with a third expert annotator reviewing their annotations. The final dataset consists of 16,914 tweets, with 3,383 instances targeting gender minorities (*sexism*), 1,972 labeled as *racism*, and 11,559 tweets neither sexist nor racist[5].

Davidson: This dataset contains 24,783 tweets[6] manually rated with three labels including *hate speech*, *offensive*, and *neither* (Davidson et al., 2017). The dataset was manually labelled by using the CrowdFlower platforms[7], where each tweet was rated by at least three annotators. The final collection only contains 5.8% of total tweets as *hate speech* and 77.4% as *offensive*, while the remaining 16.8% were labelled as *not offensive*.

Founta: This dataset collection consists of 80,000 tweets annotated with 4 mutually exclusive labels including *abusive*, *hateful*, *spam*, and *normal* (Founta et al., 2018). These tweets were gathered from the original corpus composed of 30 millions tweets was collected from 30 March 2017 to 9 April 2017. The annotation process was completed by five annotators and the final dataset is composed of 11% tweets labeled as *abusive*, 7.5% as *hateful*, 59% as *normal*, and 22.5% as *spam*.

HatEval: This dataset was used in the SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (Basile et al., 2019). It contains about 12 thousand records and its labels are hateful or not. This dataset has also been evaluated for migrants and

[3] `https://www.tensorflow.org/hub`

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

[5] We were able to retrieve only 16,488 instances (3,216 *sexism*, 1,957 *racism* and 11,315 *neither*)

[6] We only found this number in `https://github.com/t-davidson/hate-speech-and-offensive-language`

[7] Now Figure Eight `https://www.figure-eight.com/`

Dataset	Label	# Instances	Target %
Waseem (Waseem and Hovy, 2016)	**Racism**, **Sexism**, None	16,488	31.4
Davidson (Davidson et al., 2017)	**Hate Speech**, **Offensive**, Neither	24,783	83.2
Founta (Founta et al., 2018)	**Abusive**, **Hateful**, Spam, Normal	99,799	18.5
HatEval (Basile et al., 2019)	**Hateful**, Not Hateful	11,971	42.0
OLID (Zampieri et al., 2019b)	**Offensive**, Not Offensive	14,100	32.9
AbuseEval (Caselli et al., 2020)	**Abusive**, Not Abusive	14,100	20.8

Table 2: The datasets used in this paper (chronological order): labels, number of instances, and percent of records that are labeled abusive, offensive, or hateful.

misogyny.

OLID: The Offensive Language Identification Dataset (Zampieri et al., 2019a) was used in SemEval-2019 Task 6: 'OffensEval' (Zampieri et al., 2019b). It has Twitter data as the previous datasets, but it was annotated using a unique hierarchical model based on the proposed idea in (Waseem et al., 2017). We use the Offensive and Not Offensive labeled data, where about 30% of the records are labeled as Offensive.

AbuseEval: Caselli et al. (2020) created a new corpus by re-annotating OLID in order to model abusive language, seen as a correlated but independent phenomenon from offensive language. The annotation of abusiveness is carried out by three annotators at a coarse-grained, binary level (i.e., *abusive* vs. *not abusive*), and at a finer grain with the further distinction between *implicit* and *explicit* abusive language. Even though, as expected, there is overlap between offensive and abusive comments, a surprising number of instances labeled 'Offensive' in OLID were marked as 'Not Abusive' in AbuseEval.

4.2 Results and Discussion

In our experiments, we train on the training set of each of the six datasets in Table 2 and test on all of the test sets as well as the Immigrant and Misogyny test sets of HatEval, denoted as 'HatEval Mig' and 'HatEval Mis' respectively, for a total of eight test sets.

Additionally, we run the experiments with each model a total of five times and present the average result. In summary, the results presented in this section are based on 720 experiments (3 models × 6 train sets × 8 test sets × 5 runs). About the variance of the results, the average standard deviation we observe is under 0.02 with very few exceptions: for example, AbuseEval results' standard deviation has an average of 0.03.

Results for all our experiments are shown in Table 3. In this Table, we show the F1 macro-averaged results for our two models, HurtBERT-Enc (Encodings, see Figure 1) and HurtBERT-Emb (Embeddings, see Figure 2) versus the BERT baseline (refer to Section 3 for a description of all models).

Starting with in-dataset experiments (shaded gray in Table 3), the results indicate that HurtBERT performs better than the baseline on 4 out of 6 datasets, namely AbuseEval, HatEval, OLID, and Waseem. In all four cases, HurtBERT-Emb is doing the best. The improvement in F1-macro is small in some cases (e.g., for Waseem, HurtBERT-Emb has 0.838 versus 0.836 for the baseline) and larger in others (e.g., for HatEval, HurtBERT-Emb has 0.562 versus 0.533 for the baseline).

As expected, based on previous studies, the vast majority of our out-domain results are lower than the in-domain ones. For example, for Davidson, the in-domain performance (training and testing on Davidson) is in the 90's, while the out-domain (training on other datasets and testing on Davidson) ranges from 40's to 70's. There are some exceptions, for example, training our models on Founta and testing on OLID has better performance than when training our models on OLID (e.g. see BERT Baseline results, 0.753 for Founta-trained versus 0.739 for OLID trained). This is on par with previous work (Swamy et al., 2019): as they noted, there is similarity between these two datasets and Founta is a larger dataset (see Table 2).

When comparing our models with the baseline in the cross-dataset experiments, we observe that our two variants of HurtBERT obtain better results when fine-tuned on other datasets, in particular Davidson, OLID, and Waseem to a varying extent, while the results for the experiments with fine-tuning on HatEval are mixed. We observe some large improvements, for example, training

Train Set	AbuseEval			Davidson			Founta		
Test Set	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb
AbuseEval	.659	**.669**	**.686**	.577	**.578**	**.583**	.672	.657	.671
Davidson	.462	.444	.453	.908	.907	.907	.742	.738	**.745**
Founta	.707	**.715**	.702	.849	**.850**	**.850**	.916	.914	.913
HatEval	.579	.579	.571	.515	**.519**	**.517**	.532	**.539**	**.541**
HatEval Mig	.569	.554	.559	.533	**.542**	**.546**	.542	**.544**	**.578**
HatEval Mis	.572	**.582**	.567	.307	**.308**	.306	.341	**.355**	**.348**
OLID	.638	**.662**	**.666**	.663	**.667**	**.674**	.753	.741	.753
Waseem	.589	**.596**	.583	.629	**.636**	**.636**	.602	.600	**.612**

Train Set	HatEval			OLID			Waseem		
Test Set	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb
AbuseEval	.562	.548	.552	.663	**.666**	**.680**	.521	.520	**.541**
Davidson	.583	.547	.551	.703	**.704**	.703	.406	**.445**	**.462**
Founta	.570	.543	.554	.874	**.877**	.874	.512	**.516**	**.540**
HatEval	.533	**.553**	**.562**	.535	**.537**	**.540**	.524	.524	**.542**
HatEval Mig	.463	**.486**	**.483**	.575	.549	**.578**	.420	**.436**	**.450**
HatEval Mis	.598	**.638**	**.633**	.361	**.376**	**.371**	.588	.579	**.595**
OLID	.565	.545	.549	.739	.739	**.747**	.511	.507	**.536**
Waseem	.632	.614	.620	.632	.610	**.637**	.836	.834	**.838**

Table 3: The F1-macro results for all datasets. Shaded means in-dataset experiment. *B* stands for the baseline, *HB-Enc* stands for HurtBERT-Enc, and *HB-Emb* stands for HurtBERT-Emb. Bold indicates our model improves on the baseline; underlined indicates the best result (max). Each result is the average of five runs.

on Waseem and testing on Davidson, the F1-macro for HurtBERT is 0.445 (based on Encodings) and 0.462 (based on Embeddings) versus 0.406 for the BERT baseline. On the other hand, most results trained on Davidson are relatively close to the baseline. A possible explanation for this empirical evidence may have its roots in the different nature of the phenomena modeled by the datasets employed in our experiments. In fact, HurtLex seems to provide more informative knowledge to the model when the goal task is to detect *offensive* language (e.g., OLID) rather than *abusive* language (e.g., AbuseEval). This would make sense given that the lexical resource comes from a lexicon of words used to explicitly express the intention to hurt, while AbusEval is much more about “implicit” abuse.

We manually inspected some of the predictions of the models, with particular attention towards the instances that were misclassified by the baseline (BERT) and correctly classified by either HurtBERT-Enc or HurtBERT-Emb model. On HS data, we found many cases where swear words were present that are often used with non-offensive function, according to the classification in Pamungkas et al. (2020a). The word “b***h” in particular is ubiquitous in this subset, see for instance the following tweets:

> *Me: these shoes look scary Me to me: you’re a prison psychologist, suck it up,* ***b***h***
>
> *When my sister and her boyfriend was arguing my nephew went upstairs & said ”my mama not a* ***b***h*** *or a* ***h*e*** *so you better watch yo mouth”* 😂
>
> *Love that u used WOMEN instead of* ***b***h***

Our hypothesis is that the additional knowledge from HurtLex has a stabilizing effect on the representation of offensive terms, whereas the fully contextual embeddings of BERT tend to always understand such terms as offensive due to the sentence-level context.

Comparing our two models (see the model diagrams in Figures 1 and 2), we observe more improvements from HurtBERT-Emb (see again the results in Table 3). Over all the experiments, HurtBERT-Emb has the best (maximum) performance in 26 out of 48 experiments, versus 14 for HurtBERT-Enc out of 48 (there are a couple of ties in these numbers). When we look at the different training sets, the largest improvement overall is training on Waseem, where HurtBERT-Emb has the best performance among the three models in all 8 out of 8 experiments, versus only 3 out of 8 for HurtBERT-Enc. In other datasets, an example is training on OLID and testing on AbuseEval, the HurtBERT-Emb F1-macro is 0.680 versus 0.663 for the baseline and 0.666 for HurtBERT-Enc. Another example is training on Founta and testing on Hateval Mig: HurtBERT-Emb has 0.578 versus 0.542 for the baseline and 0.544 for HurtBERT-Enc.

This seems to be expected, that a method based on word embeddings performs better than one based on a simple, numerical encoding which represents an entire comment. HurtLex embeddings go through an LSTM and dense layer (Fig. 2), therefore, we expect this model to learn relationships among words in the context of the comments in the data. Nevertheless, there are cases where HurtBERT-Enc, does better; for example, training on AbuseEval and testing on Founta, HurtBERT-Enc has F1-macro of 0.715 vs 0.707 for the baseline and 0.702 for HurtBERT-Emb. This shows that, in some cases, a simple architecture with numerical encodings at the comment level can outperform the more sophisticated model based on embeddings.

5 Conclusions and Future Work

In this work, we explore how to combine a BERT model with features extracted from a hate speech lexicon. The lexical features are extracted based on multiple categories in the lexicon and according to how these categories are found in the data. The lexical features can be represented at the comment level as simple numerical encodings or at the word level as embeddings that aim to learn the relationships of the lexical features in the context of the data. We conduct extensive experimentation, with in-domain as well as cross-domain training. We observe that our methods improve on the BERT baseline in the large majority of the cases, with high gains in some cases. It proves our hypothesis that the additional features from lexical knowledge can improve the BERT performance, providing a domain-agnostic feature in a cross-domain setting. For our future work, we will explore different languages to take advantage of the multilingual aspect of our lexicon. We also plan to delve deeper into the study of the relationships between our models and the linguistic aspects and phenomena in the various abusive and offensive datasets.

Acknowledgments

The work of Valerio Basile, Endang W. Pamungkas and Viviana Patti is partially funded by Progetto di Ateneo/CSP 2016 (Immigrants, Hate and Prejudice in Social Media, S1618.L2.BOSC.01) and by the project “Be Positive!” (under the 2019 “Google.org Impact Challenge on Safety” call).

References

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the Arabic Twittersphere. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, pages 69–76. IEEE.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, 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*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*.

Sravan Bodapati, Spandana Gella, Kasturi Bhattacharjee, and Yaser Al-Onaizan. 2019. Neural word decomposition models for abusive language detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 135–145, Florence, Italy. Association for Computational Linguistics.

Arthur TE Capozzi, Mirko Lai, Valerio Basile, Fabio Poletto, Manuela Sanguinetti, Cristina Bosco, Viviana Patti, Giancarlo Ruffo, Cataldo Musto, Marco Polignano, et al. 2019. Computational linguistics against hate: Hate speech detection and visualization on social media in the “Contro L’Odio” project. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, Bari, Italy.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don’t be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Tuhin Chakrabarty, Kilol Gupta, and Smaranda Muresan. 2019. Pay “attention” to your context when classifying abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 70–79, Florence, Italy. Association for Computational Linguistics.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press.

Tullio De Mauro. 2016. Le parole per ferire. *Internazionale*. 27 settembre 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

EU Commission. 2016. Code of conduct on countering illegal hate speech online.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, volume 2150 of *CEUR Workshop Proceedings*, pages 1–15. CEUR-WS.org.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4).

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *International AAAI Conference on Web and Social Media*.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Anna Koufakou, Valerio Basile, and Viviana Patti. 2020. FlorUniTo@TRAC-2: Retrofitting word embeddings on an abusive lexicon for aggressive language detection. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 106–112, Marseille, France. European Language Resources Association (ELRA).

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Neural character-based composition models for abuse detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 1–10, Brussels, Belgium. Association for Computational Linguistics.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive Language Detection on Arabic Social Media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020a. Do you really want to hurt me? predicting abusive swearing in social media. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6237–6246. European Language Resources Association.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020b. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, page 102360.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy.

Demetris Paschalides, Dimosthenis Stephanidis, Andreas Andreou, Kalia Orphanou, George Pallis, Marios D. Dikaiakos, and Evangelos Markatos. 2020. Mandola: A big-data processing and visualization platform for monitoring and detecting online hate speech. *ACM Trans. Internet Technol.*, 20(2).

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*, pages 2798–2895. European Language Resources Association (ELRA).

Qinlan Shen and Carolyn Rose. 2019. The discourse of online content moderation: Investigating polarized user responses to changes in Reddit's quarantine policy. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 58–69, Florence, Italy. Association for Computational Linguistics.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Hong Kong, China.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the American Chapter of the Association for Computational Linguistics NAACL Student Research Workshop*, pages 88–93, San Diego, California.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 602–608.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. 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 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020).