Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3715–3723
Marseille, 20-25 June 2022

Offensive language detection in Hebrew: can other languages help?

Natalia Vanetik[1], Marina Litvak[1], Chaya Liebeskind[2], Omar Hmdia[1], Rizek Abu Madeghem[1]
[1]Shamoon College of Engineering,
Beer-Sheva, Israel,
[2]Jerusalem College of Technology,
Jerusalem, Israel
{natalyav, marinal}@ac.sce.ac.il,
liebchaya@gmail.com,
{omarhm, rezeqab}@ac.sce.ac.il

Abstract

Unfortunately, offensive language in social media is a common phenomenon nowadays. It harms many people and vulnerable groups. Therefore, automated detection of offensive language is in high demand and it is a serious challenge in multilingual domains. Various machine learning approaches combined with natural language techniques have been applied for this task lately. This paper contributes to this area in several aspects: (1) it introduces a new dataset of annotated Facebook comments in Hebrew; (2) it describes a case study with multiple supervised models and text representations for a task of offensive language detection in three languages, including two Semitic (Hebrew and Arabic) languages; (3) it reports evaluation results of cross-lingual and multilingual learning for detection of offensive content in Semitic languages; and (4) it discusses the limitations of these settings.

Keywords: offensive language, Semitic languages, deep learning, BERT, cross-lingual learning, multilingual learning

1. Introduction

Multiple works on automated offensive language detection show that contamination of social networks with offensive content is a new reality with serious outcomes affecting almost all of us. Moreover, it is an international phenomenon demanding multilingual solutions.

Early works on offensive language detection used unsupervised lexicon-based approaches (Tulkens et al., 2016), while later more supervised approaches—with logistic regression, naive Bayes, decision trees, random forests, and support vector machines—were proposed (Davidson et al., 2017). Most of the recent papers report on the application of deep neural networks—long short-term memory networks, recurrent neural networks, convolutional neural networks, gated recurrent units, transformers, and deep language models—frequently combined with word embeddings, for separating offensive language from legitimate texts (Zampieri et al., 2019c). In the last couple of years, transformer models like ELMo (Embeddings from Language Models) (Peters et al., 2018) and BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018a; Reimers and Gurevych, 2019) have been most popular and successful for offensive language identification (Liu et al., 2019; Ranasinghe et al., 2019).

The clear majority of the offensive detection studies deal with English, partially because most available annotated datasets contain English data. For example, SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval) was based on the Offensive Language Identification Dataset (OLID), which contains over 14,000 English tweets. The main findings of this task can be found in (Zampieri et al., 2019c).

Since social media became the most popular communication tool worldwide, people from different countries generate their content in various languages. The attention of international communities to this task emphasizes its "multilingual challenge" – many researchers contributed to this area by developing multilingual methodologies and annotated corpora in multiple languages. For example, such languages as Arabic (Mohaouchane et al., 2019), Dutch (Tulkens et al., 2016), French (Chiril et al., 2019), Turkish (Çöltekin, 2020), Danish (Sigurbergsson and Derczynski, 2020), Greek (Pitenis et al., 2020), Italian (Poletto et al., 2017), Portuguese (Fortuna et al., 2019), Slovene (Fišer et al., 2017), and Dravidian (Yasaswini et al., 2021) were explored for the task of offensive content identification.

Also, the multilingual methods and datasets for offensive language detection were proposed. Hate Speech and Offensive Content Identification (HASOC) 2019 (Mandl et al., 2019) and 2020 (Mandl et al., 2020) were dedicated to evaluating technology for finding Offensive Language and Hate Speech in multiple low-resource languages. HASOC 2019 provided Twitter posts for Hindi, German and English. HASOC 2020 has created test resources for Tamil and Malayalam in native and Latin scripts. Posts were extracted mainly from YouTube and Twitter. Both tracks have attracted much interest from over 40 research groups. In (Ranasinghe and Zampieri, 2020), authors addressed the mul-

tilinguality challenge by applying cross-lingual contextual word embeddings and transfer learning. They made predictions in low-resource languages, such as Bengali, Hindi, and Spanish.

Despite the great international effort, many low-resource languages got much less attention than others. Motivated by this shortage, we introduced the dataset containing annotated Facebook comments written in Hebrew in (Hmdia et al., 2021).

The general contribution of this work is multi-fold: (1) we introduce a new annotated dataset of Facebook comments in Hebrew, which is an extension of our previously published dataset with the dataset used in (Liebeskind and Liebeskind, 2018) (but not shared publicly in the original paper); (2) we perform and report monolingual experiments with multiple supervised models and text representations for a task of offensive language detection; (3) we perform cross-lingual and multilingual evaluations of the explored methods with Semitic languages as target languages; and, finally, (4) we demonstrate and discuss the limitations of this approach. The cross-lingual experiments are motivated by a big portion of low-resource languages in general and a lack of resources for Hebrew in particular. In our specific case, we take advantage of rich resources in Arabic, which is a similar language to Hebrew (both belong to the same – Semitic – family of languages). In a case of successful transfer learning from Arabic to Hebrew, one may use Arabic annotated sets for training systems aimed at the analysis of Hebrew texts. In case of success of a multilingual setting, one may augment data in one language with data in another language and train one joint multilingual model. To represent meaning of a text in different languages correctly, we use multilingual sentence embeddings and a multilingual pre-trained models, containing both languages.

2. Case study

Our case study aims at testing our hypothesis that *Semitic languages can be efficiently used in cross-lingual and multilingual learning* when not enough training data in a particular language and sufficient quality is available.

In particular, we seek answers to the following research questions:

RQ1: Can offensive language detection in Hebrew benefit from Arabic training data? We explore both replacement and enrichment of the training data in Hebrew with training data in Arabic.

RQ2: Is the observed (if any) effect symmetric? Do both languages affect each other similarly?

RQ3: Does the effect of Semitic languages on each other differ from the effect of other languages?

To explore these research questions, we built the following test cases:

- **Monolingual learning**, where each model is trained and tested in the same language. We tested multiple text representations and classification models. The main purpose of this setting was to evaluate the quality of the datasets and models, used across all test cases. We also compare the monolingual results for Semitic languages with the results of cross-lingual and multilingual settings, to see the relative effect (if any) when annotated data in a foreign language is involved in the training stage.
- **Cross-lingual learning** aims at checking whether *missing training data in a target language can be compensated by training a model in a foreign language*. Given two target languages (Hebrew and Arabic) and three training languages (Hebrew, Arabic, and English), we have four cross-lingual scenarios. Using Arabic/Hebrew for training a model, tested on Hebrew/Arabic, aims at exploring RQ1 and RQ2, respectively. Using English, which belongs to a different family but is a very high-resource language, for training aims at answering RQ3. We use multilingual sentence vectors, produced by multilingual BERT (mBERT) (Devlin et al., 2018a), for consistent representation of texts in different languages. We compare between test accuracy scores in cross-lingual and monolingual scenarios. We hope to get a smaller decline in models' performances when they are trained in foreign languages from the same family.
- **Multilingual learning** is performed for testing whether *one joint multilingual model can be trained using annotated samples in multiple languages*. We hope to get multilingual models with accuracy and f-measure higher than or comparable to the scores of respective monolingual models to accept our hypothesis. We use joint models pre-trained on a mix of languages with texts represented by multilingual sentence vectors produced by mBERT in this scenario. Given three training languages and two target languages (Semitic), we have six multilingual setups–training on two or three languages and testing on one (Hebrew or Arabic). Using Arabic/Hebrew for augmenting training data in Hebrew/Arabic aims at answering RQ1 and RQ2, respectively. Comparing multilingual models with and without involving English data aims at exploring RQ3.

Our approach is a supervised binary classification, where every text is classified into one of two classes, based on a trained model. Training sets are compiled from one or several (depending on the scenario) datasets, described below.

2.1. The data

We use three datasets to evaluate our approach, including two datasets in Semitic languages (the Arabic dataset presented in (Litvak et al., 2021) and the new

Hebrew dataset available at (Hmdia et al., 2021)). For both of these languages, a definition of hate speech as "*including communications of animosity or disparagement of an individual or a group on account of a group characteristic such as race, color, national origin, sex, disability, religion, or sexual orientation*" was used. If a text contains an offensive part, it is labeled as offensive.

Table 1 shows the data statistics for the three datasets, including their partition to a train and a test sets.

Hebrew Dataset

The Hebrew dataset is a combination of OLaH (Litvak et al., 2021) and the Liebeskind (Liebeskind and Liebeskind, 2018) datasets. Both are composed of Facebook comments written in Hebrew and annotated by humans. We took the entire OLaH collection of 2,000 annotated comments from particular Facebook groups,[1], the 1,489 annotated Facebook comments (after replacing six "unknown" labels to "positive" or "negative" and removing 211 non-Hebrew comments) from the Liebeskind dataset (Liebeskind et al., 2017)(Liebeskind and Nahon, 2017), and manually labeled additional 1,939 comments from the Liebeskind dataset[2], which were previously unlabeled.

The data was annotated by three Hebrew native speakers. Each comment was assigned two labels. In a case of disagreement between two annotators, the third one–controller–assigned the final label. The final dataset contains 5,217 annotated comments. The Kappa agreement between annotators is 0.82.

Arabic Dataset

We used the OLaA dataset, which we collected and introduced previously in (Litvak et al., 2021). OLaA is a collection of 9,000 annotated comments from Twitter. For retrieving relevant texts, we used a list of keywords, which are usually part of a typical offensive vocabulary, or describe domains usually containing offensive language in Arabic.

The Kappa agreement between annotators in OLaA is 0.75.

English Dataset

We used Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019b), which is a collection of 14,100 tweets (we used 13,240 annotated tweets from its training set). Authors report about 60% of agreement between two annotators. A third annotation and major voting were applied for the tweets with disagreement. OLID was used in the OffensEval: Identifying and Categorizing Offensive Language in Social Media (SemEval 2019 - Task 6) shared task and is available on GitHub (Zampieri et al., 2019a).

[1]ynet, the shadow, 0404 , ביתר ירושלים, תנועת רגבים, וואלה , חמל, דיווח ראשוני, ביביסטים

[2]comments to posts of Members of Israeli Knesset (MKs) between 2014–2016

2.2. Text Representation and Classification

We experiment with three different text representations – simple bag-of-words (BOW), character n-grams, and semantic representation as BERT sentence vectors.

The BOW model is good at representing word significance, and offensive words do tend to repeat themselves in users' posts. On other hand, this approach requires quality tokenization and token normalization to work well, which can be a challenge in languages with complex morphology, such as Arabic and Hebrew (it is known that tokenization and token normalization are very challenging in Semitic languages because prepositions that are commonly attached to nouns result in meaning ambiguity).

Character n-grams can assist in solving this problem by focusing on important (commonly occurred) parts of words.

However, neither BOW nor character n-grams address semantics. Therefore, we decided to employ BERT sentence vectors as semantic representation. We hope that a multilingual BERT model preserves meaning similarity across languages.

Our approach to text representation and classification (depicted in Figure 1) consists of the following steps:

1. Representing comments with one of the following:
   - BOW vectors with tf*idf weights, where every comment is treated as a separate document; vectors have length of 7,945, 38,991, and 19,732 for the Hebrew, Arabic, and English datasets, respectively;
   - character n-grams (further denoted by ng) for $1 \leq n \leq 3$; vectors have length of 10,185, 7,136, and 7,311 for the Hebrew, Arabic, and English datasets, respectively;
   - sentence vectors (further denoted by mem) generated by multilingual BERT model (Reimers and Gurevych, 2019); these vectors with 768 dimensions are used for all three languages.

2. Training and application of four ML supervised models (see Section 2.2), three of them (traditional models) with BOW, char n-grams, and multilingual BERT vectors in the monolingual scenario, and multilingual BERT vectors only in other scenarios.

We used three traditional ML models—RandomForest (RF) (Ho, 1995; Breiman, 2001), Logistic Regression (LR) (Walker and Duncan, 1967), and Support Vector Machine (SVM) with RBF kernel—and multilingual BERT (mBERT) (Devlin et al., 2018a).[3] The mBERT pre-trained model we used is b*ert-base-multilingual-cased* (Devlin et al., 2018b), with learning rate set to 0.00002, and batch size 1.

[3]We also experimented with RoBerta. However, its performance was much worse than mBERT, therefore it is not reported in this paper.

Table 1: Dataset statistics

Dataset	Source	Size	Len (min-max, avg)	Train	Test	Pos	Neg
Hebrew	Facebook	5,217	(1-489, 11)	80%	20%	40%	60%
Arabic	Twitter	9,000	(1-84, 17)	80%	20%	28%	72%
English	Twitter	13,240	(1-35,11)	80%	20%	33.1%	66.9%

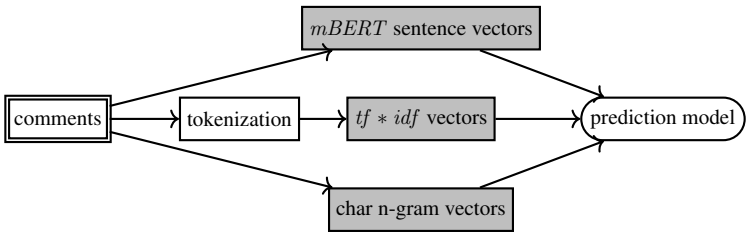

Figure 1: Offensive language detection pipeline.

3. Experiments

Our experiments aim at testing our hypothesis and answering research questions, stated at the beginning of Section 2. Below, we describe our experimental settings, and the results per test case (see Section 2), analyze common errors of automatic classification, and discuss the limitations of our study.

3.1. Setup

In monolingual experiments, we used 80% of the data for training and 20% for testing for every language. In cross-lingual experiments, we used the training data of one language (80%) and the testing data (20%) of another language. For multilingual experiments, we used the training sets of several languages combined into one training set and the test set of a target language. As result, all the methods in three setups were tested on the same test sets and therefore their results are comparable across various experiments.

3.2. Software

All baselines are implemented in sklearn (Pedregosa et al., 2011) python package. Our neural model is implemented with Keras (Chollet and others, 2015) with the TensorFlow backend (Abadi et al., 2015). Experiments were performed on a cloud server with 32GB of RAM, 150 GB of PAGE memory, an Intel Core I7-7500U 2.70 GHz CPU, and two NVIDIA GK210GL GPUs. NumPy and Pandas libraries were used for data manipulation.

3.3. Evaluation Results

Below we present results for monolingual, cross-lingual, and multilingual scenarios. Note that class distribution for the datasets used is given in Table 1.

3.3.1. Monolingual Results

Table 2 shows the evaluation results (accuracy, precision, recall, and f-measure) for all the explored models and text representations on Hebrew, Arabic, and English datasets, respectively. The best values per metric and language are colored in grey. The results of monolingual learning for Hebrew and Arabic are further compared with the respective models in cross-lingual and multilingual scenarios. The results of the English dataset are shown for providing a general picture of its quality.

As can be seen, mBERT outperforms other models in most scores for three languages. Moreover, the best recall and f-measure were consistently obtained for mBERT in all languages, meaning that mBERT does not miss as much offensive content as other models do.

The following observations were made based on comparisons between three representations: (1) character n-grams produce superior recall and f-measure compared to the BOW representation in all three models in Hebrew, and 2 out of 3 models in Arabic and English; however, its accuracy and precision are superior only in 5 and 2 cases, respectively; (2) mBERT vectors produce best f-measure in 5 out of 9 cases; and (3) using BOW vectors results in best precision scores in 7 out of 9 cases.

Note that the models with BOW/n-gram features had higher precision than the same models with mBERT vectors but much lower recall. This makes sense intuitively because certain words might be highly predictive for offensive text classification, but there might be many offensive texts that do not contain these words.

Based on our observations and the assumption that f-measure is a much more objective quality metric for the imbalanced data, we can conclude that using multilingual BERT embeddings is preferable over BOW and character n-grams as text representation with traditional ML models in monolingual data.

Being more specific, we can say that mBERT vectors dramatically help LR and SVM, but are less useful for RF, which already performs decently with BOW and n-gram features.

We also made the following observations about the models in a monolingual case: (1) mBERT transformer produces higher recall and f-measure scores for all

three languages; (2) mBERT produces the highest accuracy in Hebrew only, and it does not achieve the best precision in any language.

It is important to note that only mBERT and ML models trained on multilingual word vectors are applicable for cross-lingual and multilingual experiments. Therefore, given the superiority of mBERT and traditional ML models with mBERT vectors, we compare between performances of these models across three experiments.

3.3.2. Cross-lingual Results

Table 3 contains the evaluation scores for the cross-lingual experiments, where all models were trained on one language and tested on another—Hebrew or Arabic, respectively. As can be seen, the results are generally much worse than in mono-lingual learning scenarios, where the same language was used for training and testing.

The best results for Hebrew were obtained mostly by mBERT, trained on English. However, the best recall was observed in mBERT trained on Arabic, meaning that the Arabic-trained model recognizes offensive content in Hebrew better than English-trained, but fails in recognizing non-offensive content. Moreover, this recall score is significantly better than the respective score of the Hebrew-trained model (in the monolingual experiment), perhaps due to the larger size of the Arabic dataset.

The best scores for Arabic were achieved with different models, where LR outperformed other models in half of the cases, including the f-measure. Despite best accuracy and f-measure being produced by different models (accuracy by mBERT and f-measure by LR), both were trained on Hebrew data.

We made two conclusions from the observed results:

(1) Because most of the best results for both Semitic languages are significantly lower than the best respective monolingual results (except Ar→He recall),
our hypothesis about efficient transfer learning between similar Semitic languages in general, and in the He→Ar case (RQ2) in particular, can be rejected. However, improvement in Ar→He recall demonstrates that the lack of Hebrew training data can be compensated by Arabic data (RQ1) in a recall-oriented task, where recognizing positive samples are more important than filtering negative ones. Note that the recall of the best Hebrew model trained solely on Arabic is better than the best Hebrew model trained on Hebrew which is probably because the pre-trained multilingual model, mBERT, already contains both Arabic and Hebrew data and therefore it has a sense of the relationship between these two languages.

(2) There is a slight indication that Semitic languages are more compatible for mutual transfer learning (RQ3), however, it holds mainly for the Hebrew-to-Arabic direction. English-to-Hebrew gives better results than Arabic-to-Hebrew, but these results are worse than in a monolingual setting.

3.3.3. Multilingual Results

Table 4 contains the evaluation scores for the multilingual experiments, where the models were trained on two or three languages and tested on one—Hebrew or Arabic, respectively. The best result for every metric and target language is colored and marked in bold.

We can see that for both Semitic languages the mBERT model achieves most of the best scores.

While the best accuracy and recall for Hebrew was obtained by mBERT trained on Hebrew and Arabic, adding English samples into a training set improved precision and f-measure. The best f-measure was produced by mBERT trained on a mix of three languages. Moreover, the best multilingual scores are not far below the best monolingual ones, and there is even some improvement in recall.

The best precision for Arabic was achieved by the mBERT model trained on Hebrew and Arabic, while the best accuracy, recall, and f-measure produced by mBERT trained on English and Arabic. Similar to Hebrew, the best multilingual scores are not far below the best monolingual ones.[4]

We made the following conclusions from the observed results:

(1) Due to a marginal drop in performance of multilingual models compared with monolingual respective models, we can confirm that one joint multilingual model can be trained on training data augmented by samples from a foreign language. Also, as results support, Hebrew gains from Arabic annotated data (RQ1) more than Arabic from Hebrew (RQ2) in our task.[5]

(2) The effect of English in training scenarios does not provide any general conclusion (RQ3). In both languages, the English samples' involvement improved some of the metrics.

3.4. Error Analysis

We analyzed misclassified comments for both Semitic languages (details appear in Table 5). We used a randomly selected sample of comments that were wrongly classified for every language; the comments were evaluated by two researchers fluent in Arabic and Hebrew. We have discovered that these comments can be divided into five main classes – (1) the comments that were incorrectly annotated, to begin with; (2) the com-

[4]However, according to the Wilcoxon pairwise two-tailed non-parametric test the differences between predictions in both languages are significant, which are marked by down-arrows in Table 4.

[5]We performed an additional experiment where the number of training samples in a target language withing the mixed training set was decreased by a half. As was expected, it resulted in a drop in the mBERT performance for both languages. However, the results were much above the majority rule and testified to satisfactory performance. For example, we got Acc=0.791, F=0.755, R=0.783, and P=0.729 for a model trained on half of the Hebrew samples, and Acc=0.911, F=0.823, R=0.744, and P=0.919 for a model trained on half of the Arabic samples.

Table 2: Monolingual experiments. The evaluation results: accuracy (Acc), Precision (P), Recall (R), and F-measure (F).

	He				Ar				En			
Model	Acc	P	R	F	Acc	P	R	F	Acc	P	R	F
RF_{BOW}	0.804	0.888	0.644	0.747	0.927	0.958	0.711	0.816	0.762	0.775	0.414	0.540
RF_{ng}	0.824	0.858	0.672	0.754	0.941	0.987	0.760	0.859	0.746	0.763	0.358	0.487
RF_{mem}	0.790	0.819	0.630	0.712	0.792	0.814	0.583	0.680	0.755	0.768	0.388	0.516
LR_{BOW}	0.799	0.975	0.272	0.425	0.926	0.993	0.281	0.438	0.690	0.926	0.084	0.155
LR_{ng}	0.785	0.948	0.381	0.544	0.800	0.995	0.432	0.603	0.704	0.879	0.138	0.239
LR_{mem}	0.590	0.781	0.665	0.719	0.728	0.846	0.617	0.714	0.785	0.729	0.575	0.643
SVM_{BOW}	0.804	0.906	0.563	0.694	0.934	0.990	0.788	0.877	0.762	0.824	0.332	0.473
SVM_{ng}	0.805	0.889	0.635	0.741	0.935	0.967	0.798	0.874	0.759	0.782	0.391	0.522
SVM_{mem}	0.807	0.797	0.714	0.753	0.835	0.871	0.743	0.802	0.791	0.748	0.574	0.649
$mBERT$	0.833	0.805	0.779	0.792	0.906	0.941	0.839	0.887	0.783	0.709	0.601	0.650

Table 3: Cross-lingual experiments.

The evaluation results for Hebrew								
	Ar→He				**En→He**			
Model	Acc	P	R	F	Acc	P	R	F
RF_{mem}	0.609	0.535	0.391	0.452	0.664	0.864	0.221	0.352
LR_{mem}	0.585	0.493	0.253	0.335	0.683	**0.885**	0.267	0.411
SVM_{mem}	0.650	0.574	0.586	0.580	0.713	0.813	0.395	0.532
$mBERT$	0.412	0.449	**0.895**	0.598	**0.810**	0.835	0.695	**0.759**
The evaluation results for Arabic								
	He→Ar				**En→Ar**			
Model	Acc	P	R	F	Acc	P	R	F
RF_{mem}	0.685	0.473	0.542	0.505	0.735	0.538	0.153	0.239
LR_{mem}	0.628	0.435	**0.609**	**0.507**	0.736	**0.558**	0.169	0.259
SVM_{mem}	0.642	0.428	0.558	0.485	0.717	0.506	0.314	0.388
$mBERT$	**0.739**	0.444	0.257	0.326	0.703	0.357	0.088	0.142

Table 4: Multilingual experiments.

The evaluation results for Hebrew												
	HeAr→He				**HeEn→He**				**All→He**			
Model	Acc	P	R	F	Acc	P	R	F	Acc	P	R	F
RF_{mem}	0.770	0.832	0.563	0.671	0.777	0.832	0.577	0.681	0.769	**0.850**	0.540	0.660
LR_{mem}	0.775	0.795	0.614	0.693	0.772	0.808	0.586	0.679	0.767	0.836	0.544	0.659
SVM_{mem}	0.808	0.799	0.714	0.754	0.807	0.823	0.679	0.744	0.789	0.830	0.658	0.734
$mBERT$	**0.831**↓	0.727	**0.844**	0.781	0.823	0.819	0.735	0.775	0.822	0.783	0.788	**0.786**
The evaluation results for Arabic												
	HeAr→Ar				**ArEn→Ar**				**All→Ar**			
Model	Acc	P	R	F	Acc	P	R	F	Acc	P	R	F
RF_{mem}	0.757	0.787	0.507	0.616	0.750	0.792	0.450	0.574	0.812	0.753	0.462	0.572
LR_{mem}	0.767	0.794	0.546	0.647	0.751	0.725	0.430	0.540	0.797	0.717	0.444	0.549
SVM_{mem}	0.789	0.851	0.686	0.760	0.778	0.849	0.664	0.745	0.868	0.843	0.644	0.731
$mBERT$	0.935	**0.977**	0.737	0.840	**0.940**↓	0.944	**0.833**	**0.885**	0.926	0.956	0.770	0.853

ments where one offensive word led to the classification of a comment as offensive even though it was not; (3) the comments where these "misleading" words got an offensive context because they are frequently observed in the offensive class (probably because we created bias when used them as keywords); (4) the comments where the context (post it addresses) is missing to make a decision; and (5) all other comments where the reason for misclassification is not clear. Arabic sample did not contain any instances from categories (3) and (4).

Figure 2 contains examples of comments for the common two classes for both languages, together with their translations to English, true labels, and predicted labels by BERT. We can see that the words 'traitors' in Arabic and 'bad' in Hebrew led to labeling the posts as offensive, although their content is not negative or offensive.

3.5. Discussion

Our results on three test cases prove that mBERT is superior for most cases, especially in monolingual and

Table 5: Error classes

Language	Sample size	Wrong annotation	Word-based	Bias	Context	Other
Arabic	30	6 (20%)	1 (3%)	0	0	23 (77%)
Hebrew	26	6 (23%)	2 (8%)	5 (19%)	3 (12%)	10 (38%)

Comment (Ar/He)	Translation (En)	True label	BERT label	Error
يا حاقد ع الاسلام السياسي	O hater of political Islam	1	0	Annotation error
السيسي رءيسي وافتخر السيسي زعيمي وافتخر نعم لتعديلات الدستوريه نعم لتعديل الدستور ميعادنا ربنا يحفظ مصر وشعبها وجيشها ويجعل كيد الخونه والعملاء في نحورهم بحبك يا بلدي بحبك يا مصر نعم لتعديل دستور الاخوان	Al-Sisi is my president and I am proud Al-Sisi is my leader and I am proud, yes to amending the constitution. May God protect Egypt, its people and its army, and make the plots of **traitors** and agents to themselves. I love you, my country. I love you, Egypt. Yes, to amend the constitution of the Muslim Brotherhood	0	1	Word-based classification
קונה מעדיף בסעודיה תודה	Buyer prefers in Saudi Arabia Thank you	1	0	Annotation error
קרה בטעות חשבו שערבים רעים	Happened by mistake they thought that Arabs were **bad**	0	1	Word-based classification

Figure 2: Examples of misclassified comments for two error classes for Arabic and Hebrew, along with their translations to English.

multilingual experiments.

Also, there is evidence demonstrating an advantage of word vectors produced by mBERT as a representation over other representations in monolingual learning.

Our hypothesis that Semitic languages can be efficiently used in cross-lingual and multilingual learning was only partially confirmed for the Arabic-to-Hebrew direction. Transfer learning from Arabic to Hebrew increases recall, while extending the Hebrew training set with the data in Arabic results in almost the same accuracy score as for the monolingual setting but a higher f-measure score. This is an indication that Hebrew training data can be replaced by Arabic data if we care about recall the most, and it can be enriched with the Arabic data without the significant harm to the prediction accuracy while gaining a significant improvement in precision. Considering that the lack of resources for Hebrew was our main concern and motivation for this study, we can conclude with the recommendation for researchers who process Hebrew texts to take advantage of annotated texts in Arabic and English.

4. Conclusions and Future Work

This paper introduces a case study for offensive language detection in Semitic languages – Hebrew, Arabic – with a special focus on Hebrew, as a low-resource and morphologically-rich language. It analyzes different text representations and supervised learning methods for offensive text detection in social media. We also perform cross-lingual and multilingual experiments for testing a hypothesis about transfer learning and mutual data augmentation between two Semitic languages. Given the current results, we can recommend using transfer cross-lingual learning from Arabic to Hebrew for recall-oriented tasks and apply a joint multilingual model trained on both Arabic and Hebrew for precision-oriented tasks.

Our dataset of annotated comments written in Hebrew is publicly available and can be downloaded from GitHub (see (Hmdia et al., 2021)). In the future, we plan to explore more languages, multilingual text representations, and language models.

5. Bibliographical References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chollet, F. et al. (2015). Keras. `https://github.com/fchollet/keras`.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the

problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11(1).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018a). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018b). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Liebeskind, C. and Liebeskind, S. (2018). Identifying abusive comments in hebrew facebook. In *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*, pages 1–5. IEEE.

Liebeskind, C. and Nahon, K. (2017). Challenges in applying machine learning methods: Studying political interactions on social networks. In *Semanitic Keyword-based Search on Structured Data Sources*, pages 136–141. Springer.

Liu, P., Li, W., and Zou, L. (2019). Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 87–91.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (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.

Ranasinghe, T. and Zampieri, M. (2020). Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844.

Ranasinghe, T., Zampieri, M., and Hettiarachchi, H. (2019). Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. In *FIRE (Working Notes)*, pages 199–207.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Walker, S. H. and Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179.

6. Language Resource References

Chiril, P., Benamara, F., Moriceau, V., Coulomb-Gully, M., and Kumar, A. (2019). Multilingual and multitarget hate speech detection in tweets. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN-PFIA 2019)*, pages 351–360. ATALA.

Çöltekin, Ç. (2020). A corpus of turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184.

Fišer, D., Erjavec, T., and Ljubešić, N. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In *Proceedings of the first workshop on abusive language online*, pages 46–51.

Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., et al. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104.

Hmdia, O., Madeghem, R. A., Vanetik, N., Litvak, M., and Liebeskind, C. (2021). Hebrew dataset of offensive comments. `https://github.com/rezeq1/HebrewDataset`.

Liebeskind, C., Liebeskind, S., and HaCohen-Kerner, Y. (2017). Comment relevance classification in facebook. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 241–254. Springer.

Litvak, M., Vanetik, N., Nimer, Y., and Skout, A. (2021). Offensive language detection in semitic languages. In *1st CFP:Multimodal and Multilingual Hate Speech Detection workshop at KONVENS 2021*.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Mandl, T., Modha, S., Kumar M, A., and Chakravarthi, B. R. (2020). Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for Information Retrieval Evaluation*, pages 29–32.

Mohaouchane, H., Mourhir, A., and Nikolov, N. S. (2019). Detecting offensive language on arabic social media using deep learning. In *2019 Sixth International Conference on Social Networks Analy-*

sis, Management and Security (SNAMS), pages 466–471. IEEE.

Pitenis, Z., Zampieri, M., and Ranasinghe, T. (2020). Offensive language identification in greek. *arXiv preprint arXiv:2003.07459*.

Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., and Bosco, C. (2017). Hate speech annotation: Analysis of an italian twitter corpus. In *4th Italian Conference on Computational Linguistics, CLiC-it 2017*, volume 2006, pages 1–6. CEUR-WS.

Sigurbergsson, G. I. and Derczynski, L. (2020). Offensive language and hate speech detection for danish. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3498–3508.

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., and Daelemans, W. (2016). A dictionary-based approach to racism detection in dutch social media. *arXiv preprint arXiv:1608.08738*.

Yasaswini, K., Puranik, K., Hande, A., Priyadharshini, R., Thavareesan, S., and Chakravarthi, B. R. (2021). Iiitt@ dravidianlangtech-eacl2021: Transfer learning for offensive language detection in dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019a). Offensive Language Identification Dataset - OLID. `https://github.com/idontflow/OLID`.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019b). Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019c). 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.