Offensive language detection in Hebrew: can other languages help?

Marina Litvak*, Natalia Vanetik*, Liebeskind Chaya^, Omar Hmdia*, and Rizek Abu Madeghem*
* SCE Academic, Dept. of Software Engineering, ^ Jerusalem College of Technology, Dept. of Computer Science

SHAMOON COLLEGE OF ENGINEERING

Motivation

- offensive language in social media is a common phenomenon
- automated detection of offensive language is in high demand
- it is a serious challenge in multilingual domains
- **Hebrew is a low-resource language**

Research questions

- **RQ1**: Can offensive language detection in Hebrew benefit from Arabic training data? Or English data?
  - We explore both replacement and enrichment Hebrew training data with Arabic training data.
- **RQ2**: Is the observed (if any) effect symmetric?
  - Do both languages affect each other similarly?
- **RQ3**: Does the effect of Semitic languages one to another different from the affect of the other languages?

Our contributions

- A new annotated dataset of Facebook comments written in Hebrew
- Monolingual evaluation of multiple supervised models and text representations for a task of offensive language detection
- Cross-lingual and multilingual evaluations of the explored methods with Semitic languages as target languages

The data

Arabic
OLaA (Litvak et al., 2021)
→ 9,000 tweets
Kappa agreement 0.75
pos: 28%
neg: 72%

Hebrew
OLaH+Liebeskind
(Liebeskind et al., 2017)
new
→ 5,217 Facebook comments, additional manual labeling
Kappa agreement 0.8
pos: 40%
neg: 60%

English
OLID (Zampieri et al., 2019)
→ 14,100 tweets
13,240 used after filtering
Kappa agreement 0.6
pos: 33.1%
neg: 66.9%

שָׁלוֹם Hebrew dataset

- 5,217 comments
- taken from particular groups in Facebook:
  - ynet, the shadow, 0404 , תנועת רגבים, ביתר, ירושלים, ביביסטים, חמ"ל
- a list of Hebrew keywords was used to find offensive comments

Word in Hebrew	Translation
בושה	shame
אפס	zero
זו**נה	f***ing
זבל	trash
מחבל	terrorist
חמור	donkey (idiot)
הומו	gay
ביבי	Bibi (Netanyahu)
לפיד	Lapid (Yair)

الحكمة Arabic dataset

- 9,000 comments, written in Arabic
- a list of Arabic keywords was used to find offensive comments

Word in Arabic	Translation
يهود	Jewish
سني	Sunni
شيعي	Shiite
عربي	Arab
لقيط	bastard
ارهابي	terrorist
حمار	donkey (idiot)
دين	religions
كلب	dog

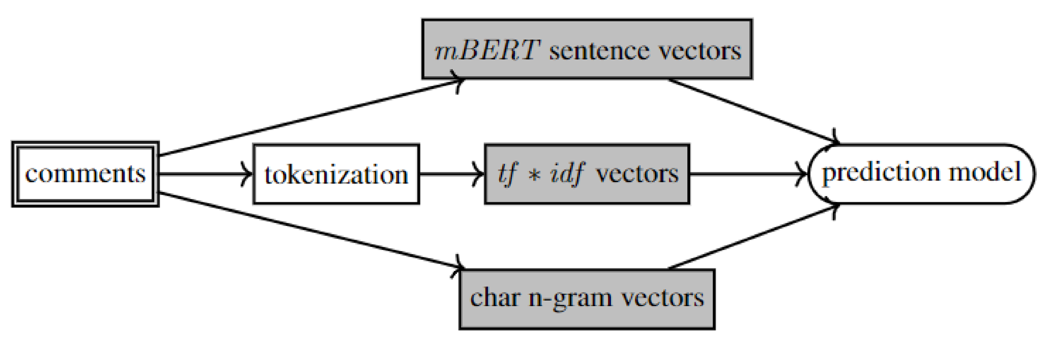
Text representation

Character n-grams $1 \le n \le 3$ → length: He 10,185; Ar 7,136; En 7,311

BOW vectors with tf-idf weights → length: He 7,945; Ar 38,991; En 19,732

Sentence vectors ml-BERT → length: 768

The pipeline



Models

1-3. RandomForest (RF),
4-6. Support Vector Machine (SVM)
7-9. Logistic Regression (LR)

applied on: char n-grams; BOW (tf*idf vectors); mBERT vectors

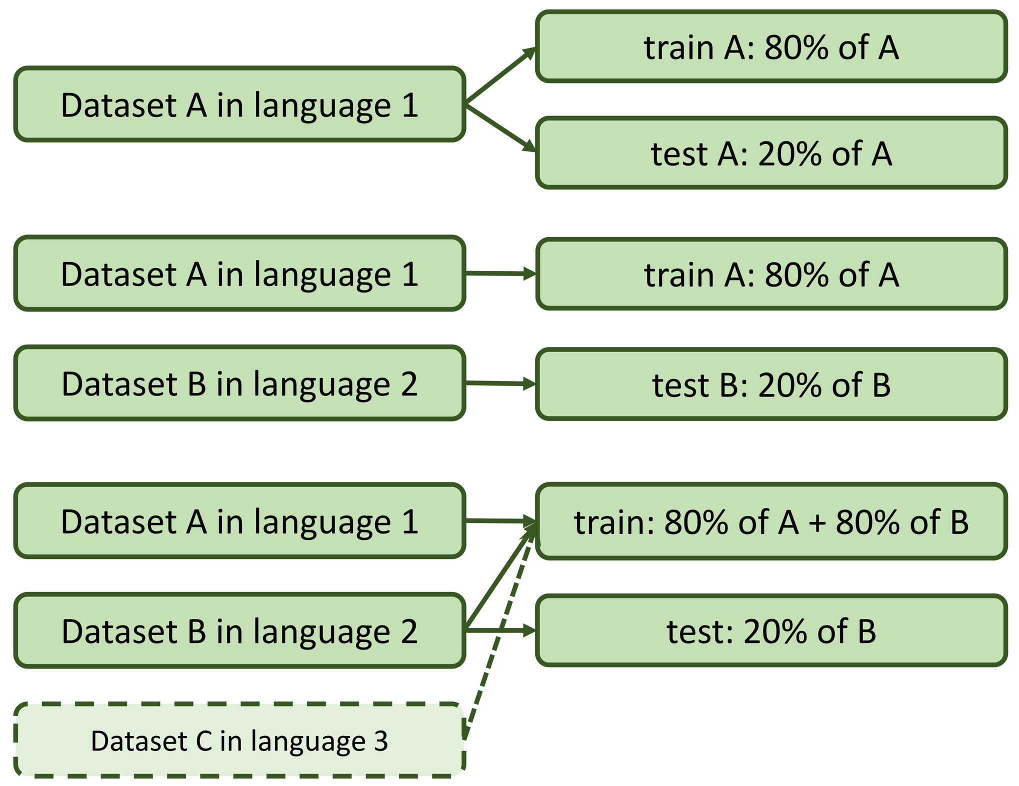
10. fine-tuned mBERT

Evaluation

Scenarios

- **Monolingual learning**: each model is trained and tested on the same language.
- **Cross-lingual learning** aims at checking whether missing training data in a target language can be compensated by training a model on a foreign language.
- **Multilingual learning** is performed for testing whether one joint multilingual model can be trained using annotated samples in multiple languages.

Train/test data split



Results

Monolingual results

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

Cross-lingual results

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

Multi-lingual results

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

Discussion

- The **mBERT model** is superior for most of cases, especially in cross-lingual and multilingual experiments.
- Weak evidence approving a possible advantage of **mBERT vectors** as a representation model in monolingual setup
- All the results achieved in the **cross-lingual** settings for Semitic languages are **significantly lower** than their monolingual results
  - except Recall in Hebrew
- Multilingual **data augmentation** performs **well** in most cases
  - extending the Hebrew training set with the data in Arabic results in the same accuracy score

Error analysis

Language	Sample size	Wrong annotation	Word-based classification	Unknown
Arabic	30	6 (20%)	1 (3.33%)	23 (76.67%)
Hebrew	30	7 (23.33%)	7 (23.33%)	`6 (53.34%)

The dataset can be downloaded from: https://github.com/ rezeq1/HebrewDataset

For questions, please contact
Dr. Chaya Liebeskind liebchaya@gmail.com
Dr. Natalia Vanetik natalyav@sce.ac.il
Dr. Marina Litvak marinal@ac.sce.ac.il