Investigating Multilingual Abusive Language Detection

Kenneth Steimel, Daniel Dakota, Yue Chen and Sandra Kübler

Department of Linguistics, Indiana University

Research Scope

- Investigate which factors have an effect on multilingual abusive language detection
- Focus on the compatibility of data and annotations
- Languages: English and German

Data Sets

English:
- Twitter hate speech dataset
- 15,715 of the 16,000 total tweets still available
- Annotation: 'racism', 'sexism', and 'none'
- 'racism' and 'sexism' mapped to 'abusive'
- 90% training data (14,143) and 10% test data (1,572)

German:
- GermEval 2018 shared task data set task 1
- Binary annotations:'offensive' or 'other'
- Use training set only

Research Questions

1. Do classifiers behave similarly across the two languages?
2. Are types of features and number of features comparable across the two languages?
3. Do over-sampling methods show consistent improvements on minority class across both languages?
4. Do the classifiers learn topic information rather than sentiment Do languages show similar effects?

Methodology

- Classifiers
  - Random Forest, XGBoost, SVM (scikit-learn)
  - Neural network architectures (keras)
- Features
  - Usupervised YASS stemmer
  - Dependency features: Dependent, head, label triples
- Sampling: Imbalanced-learn sampling suite
- Oversampling: SMOTE, Borderline SMOTE, SVMSMOTE, ADASYN
- Undersampling: Edited Nearest Neighbors, one sided selection

Results: Classifiers

	English			German		
Classifier	Prec	Rec	F-Score	Prec	Rec	F-Score
majority class	34.22	50.00	40.63	32.90	50.00	39.69
RF	80.67	74.17	76.08	66.00	66.50	66.50
XGBoost	83.46	78.80	80.49	68.50	60.00	59.50
SVM	82.11	66.58	68.20	74.41	70.97	72.01
NN	34.22	50.00	40.63	32.90	50.00	39.69

Results: Topic Modeling

	Abusive		Non-Abusive		
Language	Precision	Recall	Precision	Recall	F-score
English	33.98	53.23	70.82	52.32	52.59
German	36.97	51.46	68.32	54.41	51.80

Results: Feature Selection for English

		Overall			Abusive		
IG threshold	Num. IG features	Prec	Rec	F	Prec	Rec	F
0.000075	2660	79.38	72.62	74.49	77.95	52.02	62.39
0.00005	4232	80.29	0.74	0.76	78.95	54.44	64.44
0.000025	9305	80.72	74.27	76.18	79.59	55.04	65.08
0.00001	24350	82.26	76.48	**78.36**	81.21	59.27	**68.53**
0.0000075	33187	82.64	75.99	78.04	82.42	57.66	67.85
0.000005	60000	83.06	75.10	77.33	84.00	55.04	66.50
–	all features	81.87	66.18	67.70	87.31	34.68	49.64
–	only char n-grams	82.11	66.58	68.20	87.56	35.48	50.50

Results: Feature Selection for German

		Overall			Abusive		
IG threshold	Num. IG features	Prec	Rec	F	Prec	Rec	F
0.005	266	66.18	58.76	58.02	61.97	25.73	36.36
0.003	788	67.59	62.79	63.30	62.14	37.43	46.72
0.0014	6 404	68.70	66.23	66.92	61.65	47.95	53.95
0.0011	9 690	69.68	67.40	**68.10**	62.77	50.29	**55.84**
0.0008	16 791	70.21	65.71	66.56	65.49	43.27	52.11
0.0004	48 014	72.84	68.93	**69.95**	68.55	49.71	**57.63**
0.0002	69 541	75.21	72.28	**73.26**	70.80	56.73	**62.99**
0.0001	101 605	74.92	72.13	73.07	70.29	56.73	62.78
–	all features	74.71	71.13	72.20	70.77	53.80	61.13
–	only char n-grams	74.41	70.97	72.01	70.23	53.80	60.93

Results: Sampling for English

	Abusive		Non-Abusive		
Sampling method	Precision	Recall	Precision	Recall	F-score
No sampling	85.49	43.95	78.89	**96.56**	72.45
SMOTE	63.21	**76.21**	**87.89**	79.55	76.31
Borderline SMOTE	62.23	73.99	86.92	79.65	75.48
SVM SMOTE	62.46	73.79	86.82	79.55	75.34
ADASYN	61.19	74.40	86.89	78.25	74.75
Edit nearest neighbors	81.77	57.86	82.88	94.05	**77.94**
One sided selection	**85.60**	44.35	79.01	**96.56**	72.67

Results: Sampling for German

	Abusive		Non-Abusive		
Sampling method	Precision	Recall	Precision	Recall	F-score
No sampling	**70.80**	56.73	79.61	**87.84**	**73.26**
SMOTE	58.17	52.05	76.37	80.55	66.67
Borderline SMOTE	60.26	54.97	77.62	81.16	68.42
SVM SMOTE	60.26	54.97	77.62	81.16	68.42
ADASYN	57.32	52.63	76.38	79.64	66.43
Edit nearest neighbors	56.81	**70.76**	**82.58**	72.04	69.98
One sided selection	69.57	56.14	79.28	87.23	72.60

Discussion

- Need different classifiers
- Useful features:
  - Stems & dependencies helpful for English but not German
  - German: less than half of features; English: only 4.5% of features
- Sampling: Undersampling: effective for English only; all features better than sampling
- Topics: No meaningful overlap between topics and non-abusive/abusive language

Conclusions

- Best approach for the two languages differ largely
- Multilingual approaches to abusive language detection need more work