Detecting Hate Speech in Tweets

LING 573 Deliverable 4
Team: PlaceboAffect
(Mohamed Elkamhawy, Karl Haraldsson, Alex Maris, Nora Miao)

Our Team: **PlaceboAffect**

Mohamed Elkamhawy
mohame@uw.edu

Karl Haraldsson
kharalds@uw.edu

Alex Maris
alexmar@uw.edu

Nora Miao
norah98@uw.edu

The Shared Task: Detecting Hate Speech

- **SemEval 2019 Task 5**: Multilingual detection of hate speech against immigrants and women in Twitter (HatEval)
- **Language**: English (primary) and Spanish (adaptation)
- **Recognition type**: binary classification (hate speech or non-hate speech)
- **Affect type**: attitude
- **Target:** aspect-specific
- **Genre**: tweets
- **Modality**: text

The Data

- Dataset requested via http://hatespeech.di.unito.it/hateval.html
- Collected from July to November 2017 for **women-targeted tweets** and July to September 2018 for **immigrant-targeted tweets**
- **English** language dataset: **13,000 tweets** (9,000 training, 1,000 development, and 3,000 testing)
- **Spanish** language dataset: **6,600 tweets** (4,500 training, 500 development, and 1,600 testing)

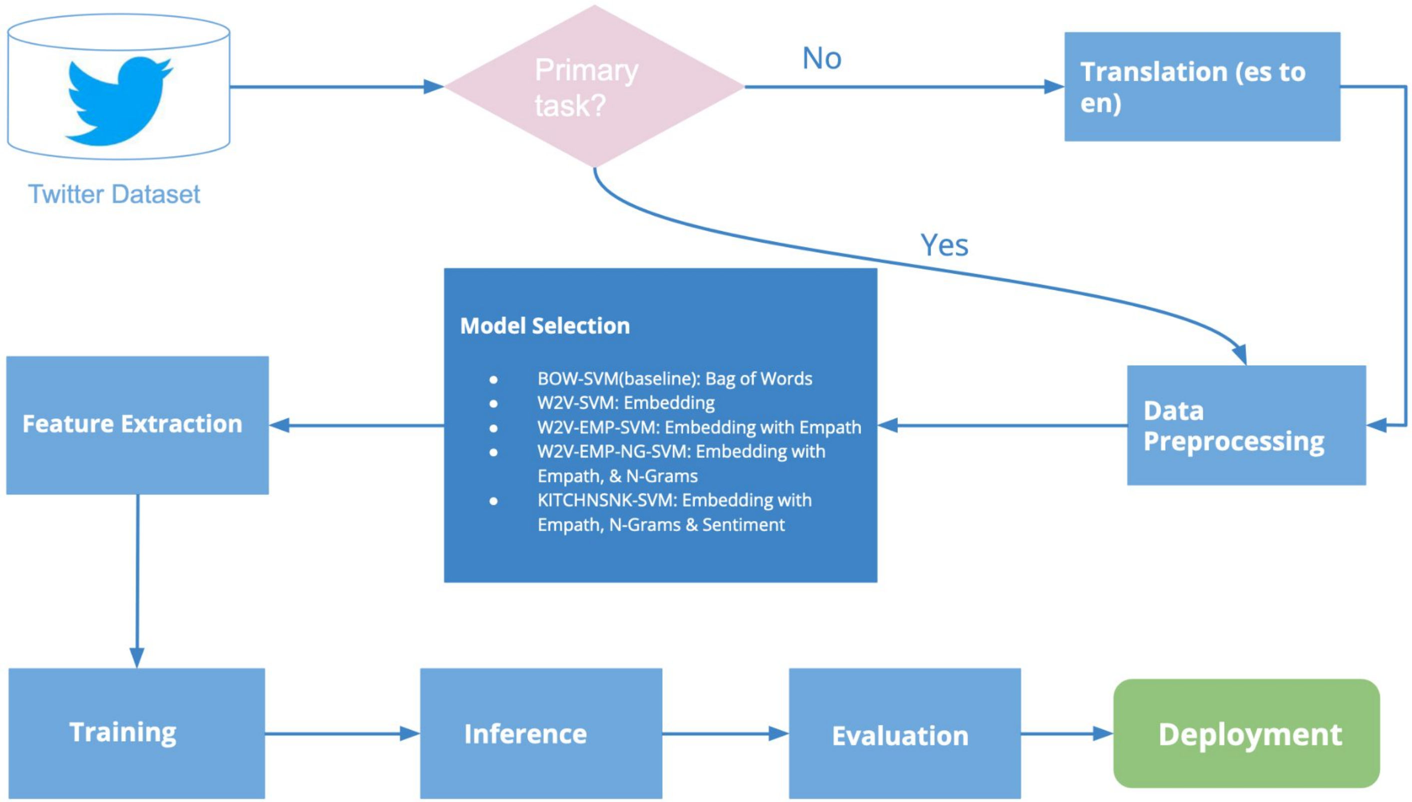
See Appendix A for more details regarding annotation guidelines for distinguishing hate speech

Examples

@[redacted] Hurray, saving us $$$ in so many ways @potus @realDonaldTrump #LockThemUp #BuildTheWall #EndDACA #BoycottNFL #BoycottNike **[HATEFUL]**

@[redacted] Orban in Brussels: European leaders are ignoring the will of the people, they do not want migrants https://t.co/NeYFyqvYIX **[NOT HATEFUL]**

System Architecture



Approach

	Preprocessing	Feature Extraction	Training	Inference	Evaluation
Initial System	● Tokenize (`nltk` tweet tokenizer) ● Lemmatize tokens w/`WordNet` lemmatizer	● Apply CBOW `word2vec` to produce average tweet embeddings ● Concatenate lexical information from `empath` library	● `SVC` from sklearn ● Grid search for best hyperparameters using dev set ● Select and save model with best `f1_macro` score	● Use best model from training step (or load model if only performing inference) ● Feed dev instances into `model.predict()` ● Save predictions to output/	● Compare predictions to golden labels, applying sklearn's accuracy measure as well as macro-averaged versions of precision, recall and f1. ● Output results to txt file.
Enhanced System	● Convert emojis to text (`emoji.demojize()`) ● Handle negation	● Concatenate sentiment score from `vaderSentiment`	● Tested alternative classification algorithms, e.g., XGBoost, Naive Bayes, Random Forest,etc		
Adaptation	● `deep-translator` to translate Spanish tweets into English		● concatenate english and spanish (translated) training sets		
Inputs & Outputs	I: .csv file with tweets O: arrays of tweets & labels	I: array of tweets & labels O: train and dev vectors	I: training vectors O: model object, model file	I: model object, unlabeled instances O: predictions	I: predictions O: results file
Module(s)	`features.preprocess`	`features.extract_features`	`modeling.classifier`	`modeling.classifier`	`main.evaluate`

Results (Primary)

Our model substantially outperformed the SemEval 2019 Task 5 Baseline for the Subtask. Also it outperformed our own bag of words baseline model F1-macro (BOW-SVM) by 0.10.

Model	*Acc*	*Prec*	*Rec*	*F1-Macro*
W2V-SVM (alpha)	**0.52**	0.54	0.54	**0.52**
W2V-EMP-SVM (beta)	0.50	0.54	0.55	0.49
W2V-EMP-NG-SVM (gamma)	**0.52**	**0.56**	**0.60**	0.49
KITCHNSNK-SVM (delta)	0.49	0.55	0.58	0.46
KITCHNSNK-XGBoost (delta)	0.51	**0.56**	**0.60**	0.49
BOW-SVM (D4 baseline)	0.47	0.53	0.56	0.42
tf-idf SVC (Shared Task *Baseline*)	**-**	**-**	**-**	0.45

Results (Adaptation)

Our model substantially outperformed the SemEval 2019 Task 5 Baseline for the Subtask.

Model	*Acc*	*Prec*	*Rec*	*F1-Macro*
W2V-SVM (alpha)	0.66	0.64	0.65	0.64
W2V-EMP-SVM (beta)	0.70	0.65	**0.77**	0.64
W2V-EMP-NG-SVM (gamma)	0.72	0.68	0.76	0.68
KITCHNSNK-SVM (delta)	**0.73**	0.70	0.74	**0.71**
KITCHNSNK-XGBoost (delta)	**0.73**	0.70	0.73	**0.71**
BOW-SVM (D4 baseline)	0.72	**0.71**	0.71	0.71
tf-idf SVC (Shared Task *Baseline*)	**-**	**-**	**-**	0.70

Issues and successes

- Adaptation task methodology: translating Spanish to English
    - (+) Reuse English lexical features from primary task
    - (+) Ability to concatenate primary (English) and adaptation (Spanish) training data
    - (-) Performance depends on quality of translation
    - (-) Lose cultural context with translation, and hate speech is often correlated with language-specific slang

Issues and successes

- Better performance on dev set compared to test set. Overfit to dev set?

*Dataset (**Primary** task)*	*Model*	*F1-Macro*
Dev	W2V-SVM (alpha)	0.65
Test	W2V-SVM (alpha)	0.52

*Dataset (**Adaptation** task)*	*Model*	*F1-Macro*
Dev	KITCHNSNK-SVM (delta)	0.76
Test	KITCHNSNK-SVM (delta)	0.71

- Inefficiencies in training, e.g. speed

- Communication norms established early, well-defined task breakdowns

Related Readings

- Doris Chinedu Asogwa, Chiamaka Ijeoma Chukwuneke, CC Ngene, and GN Anigbogu. 2022. Hate speech classification using SVM and Naive BAYES. arXiv preprint arXiv:2204.07057.
- 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.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.
- Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, volume 2150, pages 214–228.
- Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657.

Thank you.

Appendices

Appendix A: Annotation Guidelines

HS against immigrants may include:

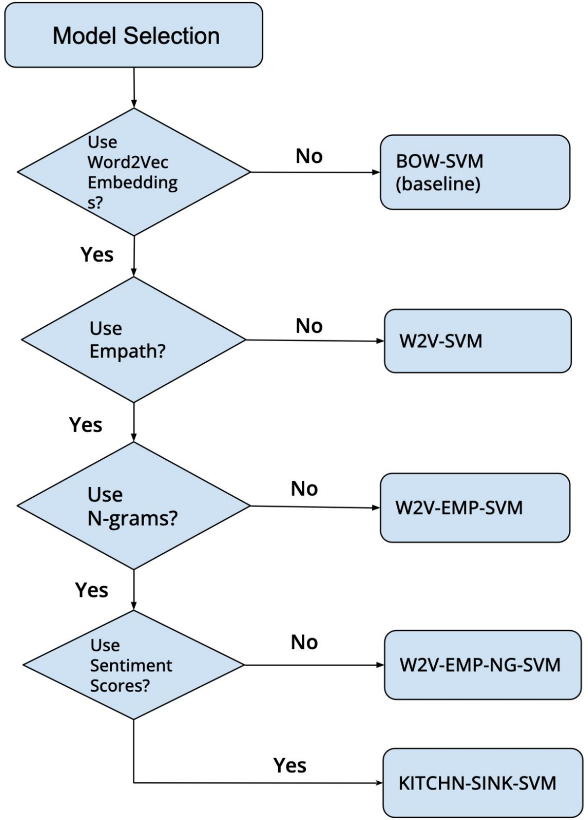
- insults, threats, denigrating or hateful expressions
- incitement to hatred, violence or violation of rights to individuals or groups perceived as different for somatic traits (e.g. skin color), origin, cultural traits, language, etc.
- presumed association of origin/ethnicity with cognitive abilities, propensity to crime, laziness or other vices
- references to the alleged inferiority (or superiority) of some ethnic groups with respect to others
- delegitimation of social position or credibility based on origin/ethnicity
- references to certain backgrounds/ethnicities as a threat to the national security or welfare or as competitors in the distribution of government resources
- dehumanization or association with animals or entities considered inferior

Tweets had to meet BOTH of the following criteria:

1. the tweet content MUST have I**MMIGRANTS/REFUGEES as main TARGET,** or even a single individual, but considered for his/her membership in that category (and NOT for the individual characteristics)
2. we must deal with a message that spreads, incites, promotes or justifies **HATRED OR VIOLENCE TOWARDS THE TARGET, or a message that aims at dehumanizing, hurting or intimidating the target**

See https://github.com/msang/hateval/blob/master/annotation_guidelines.md

Appendix B: Model Selection



Appendix C: Issues and successes

- High amount of false negatives for the W2V approaches
- Improvement in recall when using more lexical features (for English evaltest)

English Evaltest

Model	*TP*	*TN*	*FP*	*FN*
BOW-SVM	1134	277	1463	126
W2V-SVM	845	713	1027	415
W2V-EMP-SVM	979	528	1212	281
W2V-EMP-NG-SVM	1090	460	1280	170
KITCHNSNK-SVM	1120	361	1379	140

Spanish Evaltest

Model	*TP*	*TN*	*FP*	*FN*
BOW-SVM	423	727	213	237
W2V-SVM	329	727	213	331
W2V-EMP-SVM	218	908	32	442
W2V-EMP-NG-SVM	277	877	63	383
KITCHNSNK-SVM	349	821	119	311

Appendix C: Issues and successes (**evaltest primary**)

- Scores of words containing ‘stupid’ for the BOW approach

	Negative	Positive
Model predictions	**2**	**13**
True labels	13	2

- Scores of words containing ‘stupid’ for the W2V-SVM approach

	Negative	Positive
Model predictions	**7**	**8**
True labels	13	2

- Scores of words containing ‘stupid’ for the W2V-EMP-SVM approach

	Negative	Positive
Model predictions	**7**	**8**
True labels	13	2

Appendix C: Issues and successes (**evaltest adaptation**)

Tweets containing “deportation”

Spanish evaltest:

	Negative (non-HS)	Positive (HS)
BOW predictions	3	4
W2V predictions	6	1
EMP predictions	6	1
NG predictions	5	2
KS predictions	3	4
True labels	**0**	**7**

Appendix C: Issues and successes

- Train vs. dev vs. test set distributions

*Dataset (**Primary** task)*	*# documents*	*% hate speech*
Train	9000	42.0%
Dev	1000	42.7%
Test	3000	42%

*Dataset (**Adaptation** task)*	*# documents*	*% hate speech*
Train*	13500	41.8%
Dev	500	44.4%
Test	1600	41.3%

*adaptation training set consisted of the original adaptation training data and the primary training data

Appendix C: Issues and successes (**evaltest primary**)

- TP tweets had the highest average empath ‘hate’ score, then TN and FP, and last FN

	TP	TN	FP	FN
Average score of ‘hate’	**0.00462**	0.00341	0.00324	0.00318

- ‘swearing_terms’ scores had the largest difference between true labeled hate speech (TP, 0.0601) and non hate speech (FP, 0.0187)

- However, the model classified high scoring ‘swearing_terms’ tweets as positive for hate speech

	TP	TN	FP	FN
Average score of ‘swearing_terms’	**0.0601**	0.0187	**0.0346**	0.0152

Appendix C: Issues and successes (**evaltest adaptation**)

- FP tweets had the highest average empath ‘hate’ score, then FN and TP, and last TN

	TP	TN	FP	FN
Average score of ‘hate’	0.00275	0.00253	**0.00486**	0.00441

- ‘swearing_terms’ scores had the largest difference between true labeled hate speech (TP, 0.1078) and non hate speech (FP, 0.0701)

- However, the model classified high scoring ‘swearing_terms’ tweets as positive for hate speech

	TP	TN	FP	FN
Average score of ‘swearing_terms’	**0.1078**	0.0090	**0.0701**	0.0116