Detecting Hate Speech in Tweets

LING 573 Deliverable 2
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)
- **Recognition type**: binary classification (hate speech or non-hate speech)
- **Affect type**: attitude
- **Target:** aspect-specific
- **Genre**: tweets
- **Modality**: text
- **Language**: English (primary) and Spanish (adaptation)

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

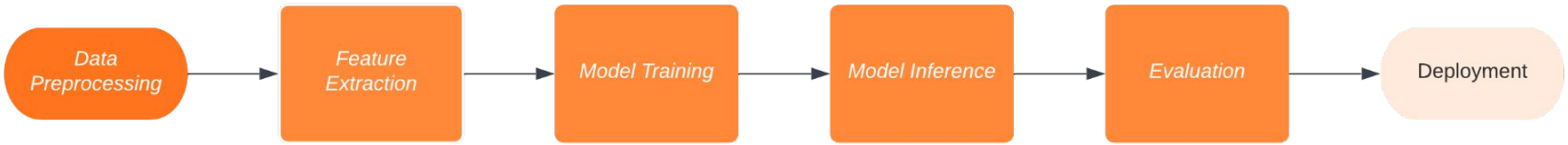
Examples

@[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/NeYFyqvYlX **[NOT HATEFUL]**

System Architecture

- Three models are evaluated:
  - The baseline model, BOW-SVM, uses a Bag-of-Words representation for the features and a binary SVM classifier
  - Our first proposed model, W2V-SVM, also uses a binary SVM classifier but relies on embeddings derived using Word2Vec
  - Our second proposed model, W2V-EMP-SVM, uses additional lexical features derived using the empath package



Approach

	Preprocessing	Feature Extraction	Training	Inference	Evaluation
Description	• Tokenize (`nltk` tweet tokenizer) • Lemmatize tokens w/`WordNet` lemmatizer • Remove English stopwords (`nltk` list)	• Apply CBOW `word2vec` to produce average tweet embeddings • [SVM-EMP-W2V Only] Concatenate lexical information from `empath` library	• `SVC` from sklearn • Grid search for best hyperparameters using 5-fold cross-validation on the *training* set only (`GridSearchCV`) • 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.
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

Our model substantially outperformed the SemEval 2019 Task 5 Baseline for the Subtask, but it underperformed our own bag of words baseline model F1-macro (BOW-SVM) by 0.06.

Model	*Acc*	*Prec*	*Rec*	*F1-Macro*
W2V-SVM	0.67	0.62	0.71	0.60
W2V-EMP-SVM	0.70	0.66	0.72	0.65
BOW-SVM (Our Baseline)	**0.73**	**0.72**	**0.73**	**0.72**
tf-idf SVC* (Shared Task *Baseline*)	**-**	**-**	**-**	**0.45**

**Note that the Basile et al 2019 baseline was calculated on the test set, not the development set. Thus, this baseline is included for reference only and cannot be used in formal evaluation until D4.*

Issues and successes

- High amount of false negatives for the W2V approaches
- Slight improvement in recall when using EMP

Model	*TP*	*TN*	*FP*	*FN*
BOW-SVM	281	452	121	146
W2V-SVM	131	537	36	296
W2V-EMP-SVM	174	522	51	253

Issues and successes

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

	Negative	Positive
Model predictions	**5**	**12**
True labels	3	14

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

	Negative	Positive
Model predictions	**8**	**9**
True labels	3	14

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

	Negative	Positive
Model predictions	**10**	**7**
True labels	3	14

Issues and **successes**

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

	TP	TN	FP	FN
Average score of ‘hate’	**0.00365**	0.00185	0.00262	0.00234

- ‘swearing_terms’ scores had the largest difference between true labeled hate speech (0.01975) and non hate speech (0.0060)

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

	TP	TN	FP	FN
Average score of ‘swearing_terms’	**0.04274**	0.00206	**0.04576**	0.00393

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.

What’s Next

- Adjustments to preprocessing to curtail information loss?
  - Further adjustments to handling emoji, hashtags, accounts, and misspellings?
- Expanding the predictor features?
  - Concatenating embedding vectors with BOW vector?
- New architectures altogether?
  - Fine-tuning LLM text classifier w/HuggingFace

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