Multilingual Detection of Hate

Ben Cote, Madhav Kashyap, Lindsay Skinner, Keaton Strawn, Allan Tsai

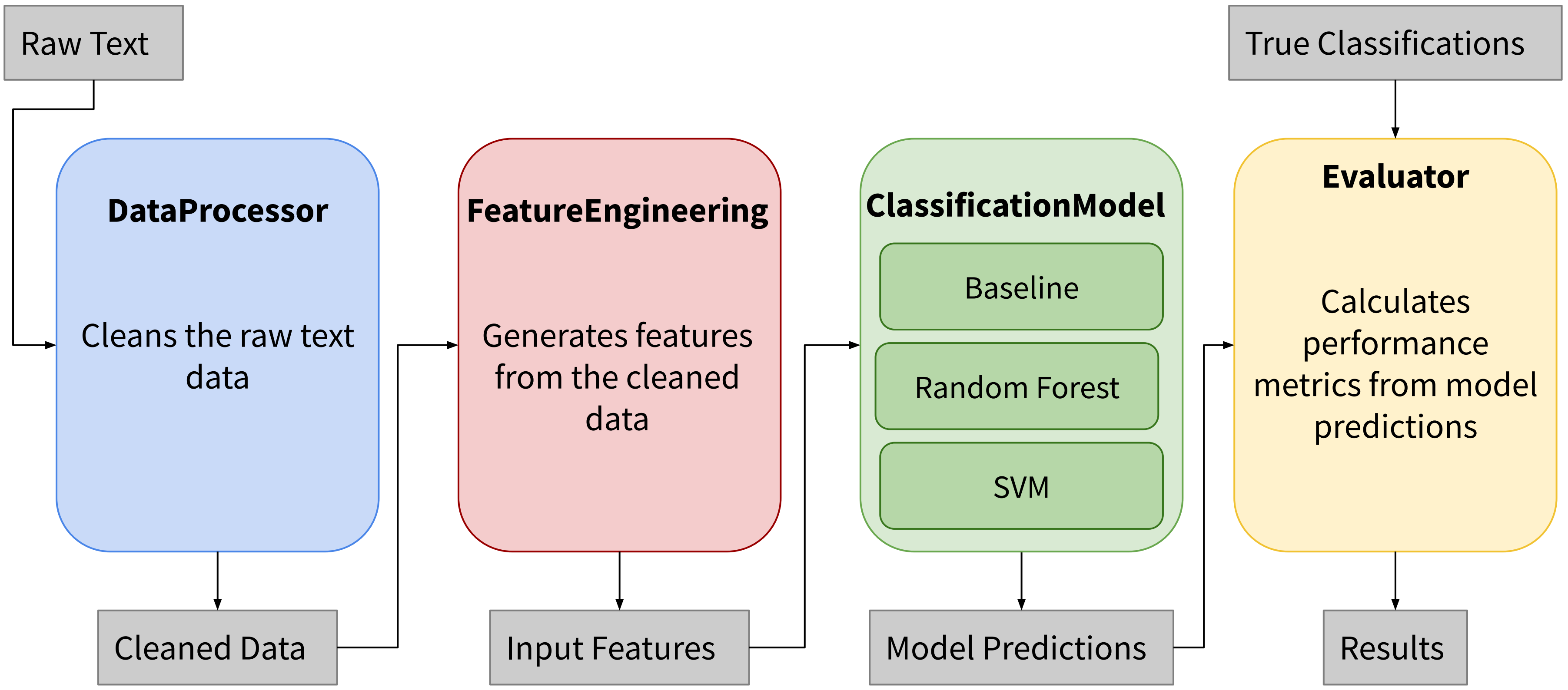
Task Background – SemEval 2019: Task 5

- Multilingual Detection of Hate against Immigrants and Women in Twitter data (English for D1-3, Spanish for D4 adaptation), two subtasks: Task A and Task B
- Data is pre-labeled tweets, classified with either a 1 or a 0 on three metrics: Hate Speech, Personal Target, and Aggression
- Hate Speech (Task A) refers to whether tweet disparages a person or a group on the basis of characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion, etc.
- Personal Target (Task B) refers to hate speech tweets only, reflecting whether they are targeting a group generally or a specific person/user
- Aggression (Task B) reflects whether or not a hate speech tweet is aggressive

Prior Work

- SVM + Universal Sentence Encoder sentence-level embeddings
- Bidirectional Gated Recurrent Units (BiGRUs) + fastText word embeddings
- Multiple Choice CNN + BERTweet embeddings
- LSTM + standard GloVe embeddings + data augmentation with paraphrasing tools
- + many more approaches for both tasks and both languages

Our System



Raw Text
True Classifications
DataProcessor
Cleans the raw text data
FeatureEngineering
Generates features from the cleaned data
ClassificationModel
Baseline
Random Forest
SVM
Evaluator
Calculates performance metrics from model predictions
Cleaned Data
Input Features
Model Predictions
Results

DataProcessor

- Separates URLs and hashtags from the text
- Replaces emojis with English descriptions
- Calculates and adds a feature indicating what percentage of the text is capitalized
- Lower-cases the cleaned text
- Generates counts for special punctuation symbols
- Removes all punctuation from the cleaned text
- Stores twitter user IDs in a feature and replaces them with the string 'user'
- Replaces typos with best guesses of correctly spelled words from the python spellchecker library

INPUT

Raw text: "Hurray, saving us $$$ in so many ways @potus @realDonaldTrump #LockThemUp #BuildTheWall #EndDACA #BoycottNFL #BoycottNike"

OUTPUT

Cleaned text: "hurray saving us in so many ways user user"

Hashtags: [#LockThemUp #BuildTheWall #EndDACA #BoycottNFL #BoycottNike]

User IDs: [@potus, @realDonaldTrump]

Percent capitalized: 0.029412

$ count: 3

FeatureEngineering

- Normalized counts for punctuation
- Normalized NRC lexicon Counts: binary classification of words across eight emotional dimensions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) and a positive dimension and negative dimension
- Aggregated GloVe twitter embeddings (d=25) (sentence-level average of word embeddings)
- Sentence embeddings from Google's Universal Sentence Encoder (d=512)
- Aggregated BERTweet embeddings (d=768) (sentence-level average of words)
- Slang dictionary: uses data from the SlangSD to label slang words with sentiment strength from -2 to 2, where -2=strongly negative, -1=negative, 0=neutral, 1=positive, 2=strongly positive. Scores are aggregated across all the slang words in a tweet

INPUT
Cleaned text: "hurray saving us in so many ways user user"
Hashtags: [#LockThemUp #BuildTheWall #EndDACA #BoycottNFL #BoycottNike]
User IDs: [@potus, @realDonaldTrump]
Percent capitalized: 0.029412
$ count: 3

OUTPUT
! count normalized: 0.385523
? count normalized: 0.428634
$ count normalized: 1.0
*** count normalized:** 0.46831
Slang score: 0
Universal sentence encoder embeddings: [0.045,...]
BERTweet embeddings: [-0.015,...]
Aggregate GloVe embeddings: [0.303,...]

ClassificationModel

- Baseline
  - Predicts the most seen label for each instance
- Random Forest
  - Predictions based on number of trees, split criterion, minimum split size, maximum number of features, equal vs. balanced class weights, maximum samples per tree
  - Able to treat the classification as three separate binary tasks, or one five-class classification task
- Multi-class SVM
  - Predictions based on kernel type and number of degrees (for polynomial kernels)
  - Able to treat the classification as three separate binary tasks, or one five-class classification task
- (In-Progress: Fine-tuned BERT for Sequence Classification with raw data as input)

INPUT
Percent capitalized: 0.029412
! count normalized: 0.385523
? count normalized: 0.428634
$ count normalized: 1.0
*** count normalized:** 0.46831
Slang score: 0
Universal sentence encoder embeddings: [0.045,...]
BERTweet embeddings: [-0.015,...]
Aggregate GloVe embeddings: [0.303,...]

OUTPUT

HS (hate speech): 0
TR (targeted): 0
AG (aggressive): 0

Evaluator

Calculates F1 across all classes and for individual classes, precision for each class, recall for each class, accuracy for each class, and exact match ratio

INPUT

Predictions:

Tweet 1: HS=0, TR=0, AG=0
Tweet 2: HS=1, TR=0, AG=0
Tweet 3: HS =1, TR=0, AG=0
Tweet 4: HS=1, TR=1, AG=0
…

Actual:

Tweet 1: HS=1, TR=0, AG=0
Tweet 2: HS=1, TR=0, AG=0
Tweet 3: HS =1, TR=0, AG=0
Tweet 4: HS=1, TR=0, AG=1
…

OUTPUT

Macro F1: 0.449
EMR: 0.527
HS F1: 0.465
TR F1: 0.439
AG F1: 0.443
HS Precision: 0.551
TR Precision: 0.891
AG Precision: 0.898
HS Recall: 0.521
TR Recall: 0.898
AG Recall: 0.891
HS Accuracy: 0.577
TR Accuracy: 0.781
AG Accuracy: 0.796

Evaluation and Results: Top 5 SVM models

Kernel	Degree	Macro F1	EMR	HS F1	TR F1	AG F1
polynomial	5	0.614	0.541	0.451	0.447	0.943
linear	N/A	0.479	0.479	0.559	0.439	0.443
polynomial	2	0.466	0.511	0.515	0.439	0.443
polynomial	6	0.461	0.543	0.454	0.482	0.447
polynomial	4	0.449	0.539	0.466	0.439	0.443

D2 Results						
Kernel	**Degree**	**Macro F1**	**EMR**	**HS F1**	**TR F1**	**AG F1**
polynomial	3	0.415	0.173	0.299	0.640	0.169

Evaluation method provided by SemEval 2019: Task 5 description

Evaluation and Results: Top 5 Random Forest models

Number of Trees	Splitting Criterion	Min Split Size	Max Features per Tree	Max Samples per Tree	Macro F1	EMR	HS F1	TR F1	AG F1
500	entropy	0.2	sqrt	0.5	0.641	0.488	0.545	0.934	0.443
500	gini	0.5	log2	0.7	0.634	0.452	0.523	0.936	0.443
500	entropy	0.5	log2	0.7	0.634	0.468	0.523	0.935	0.443
1000	gini	0.5	log2	0.7	0.631	0.461	0.517	0.934	0.443
2000	entropy	0.5	sqrt	0.7	0.631	0.460	0.513	0.936	0.443

*Each of the top 5 RF models used balanced class weights

Evaluation method provided by SemEval 2019: Task 5 description

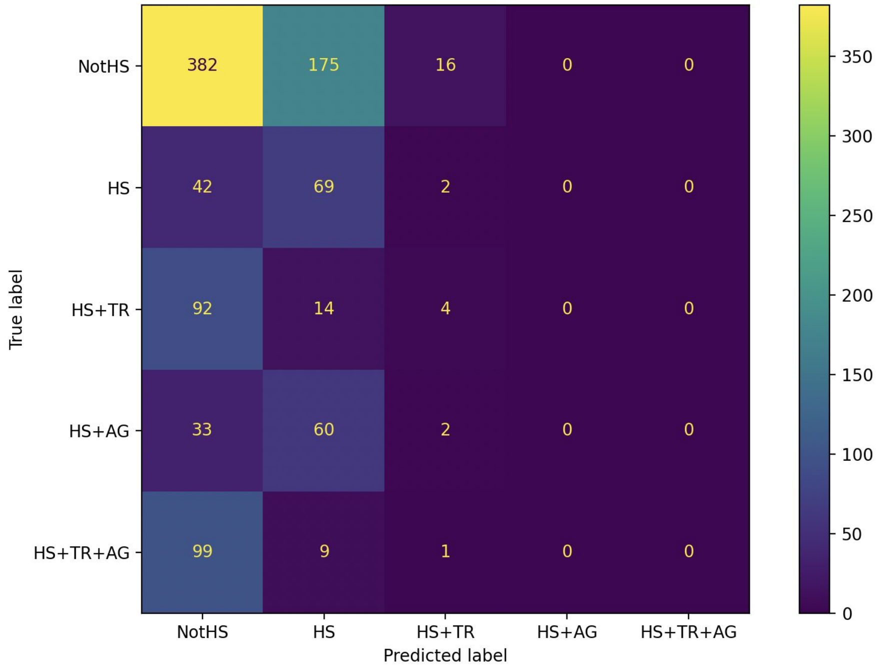
Detailed Results for Top Performing Model

Approach	n_trees	criterion	min_split	max_features	class_wts	max_samples
random_forest	500	entropy	0.2	sqrt	balanced	0.5

Random Forest Results													
Macro F1	**EMR**	**Accuracy HS**	**Accuracy TR**	**Accuracy AG**	**Precision HS**	**Precision TR**	**Precision AG**	**Recall HS**	**Recall TR**	**Recall AG**	**F1 HS**	**F1 TR**	**F1 AG**
0.641	0.488	0.578	0.766	0.796	0.558	0.389	0.898	0.550	0.898	0.389	0.545	0.934	0.443

Baseline Results													
Macro F1	**EMR**	**Accuracy HS**	**Accuracy TR**	**Accuracy AG**	**Precision HS**	**Precision TR**	**Precision AG**	**Recall HS**	**Recall TR**	**Recall AG**	**F1 HS**	**F1 TR**	**F1 AG**
0.415	NA	0.573	0.781	0.796	0.787	0.891	0.898	0.500	0.500	0.500	0.364	0.439	0.443

Detailed Results for Top Performing Model: Confusion Matrix



Next Steps

- Adapt for Spanish
  - New embeddings
  - New slang dictionary
- NRC Lexicon Extension
- Update slang dictionary (address stop words)
- Add hashtag and emoji embeddings
- Ablation tests and further parameter tuning
- Data Augmentation
- Use GPU to speed up processing (BERT)
- Add fine-tuned BERT and ensemble results, if it performs well

Color coding:

- Required additions that have not yet been started
- Additions that are nearly complete
- Nice to have additions, that will be included if we have time

References

- 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.