Multilingual Detection of Hate

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

Task Description – SemEval 2019: Task 5

Multilingual Detection of Hate against Immigrants and Women in Twitter data

Primary Task

- Data is pre-labeled tweets in English
- Binary classification (0/1) on Hate Speech, Aggression, and Target Group

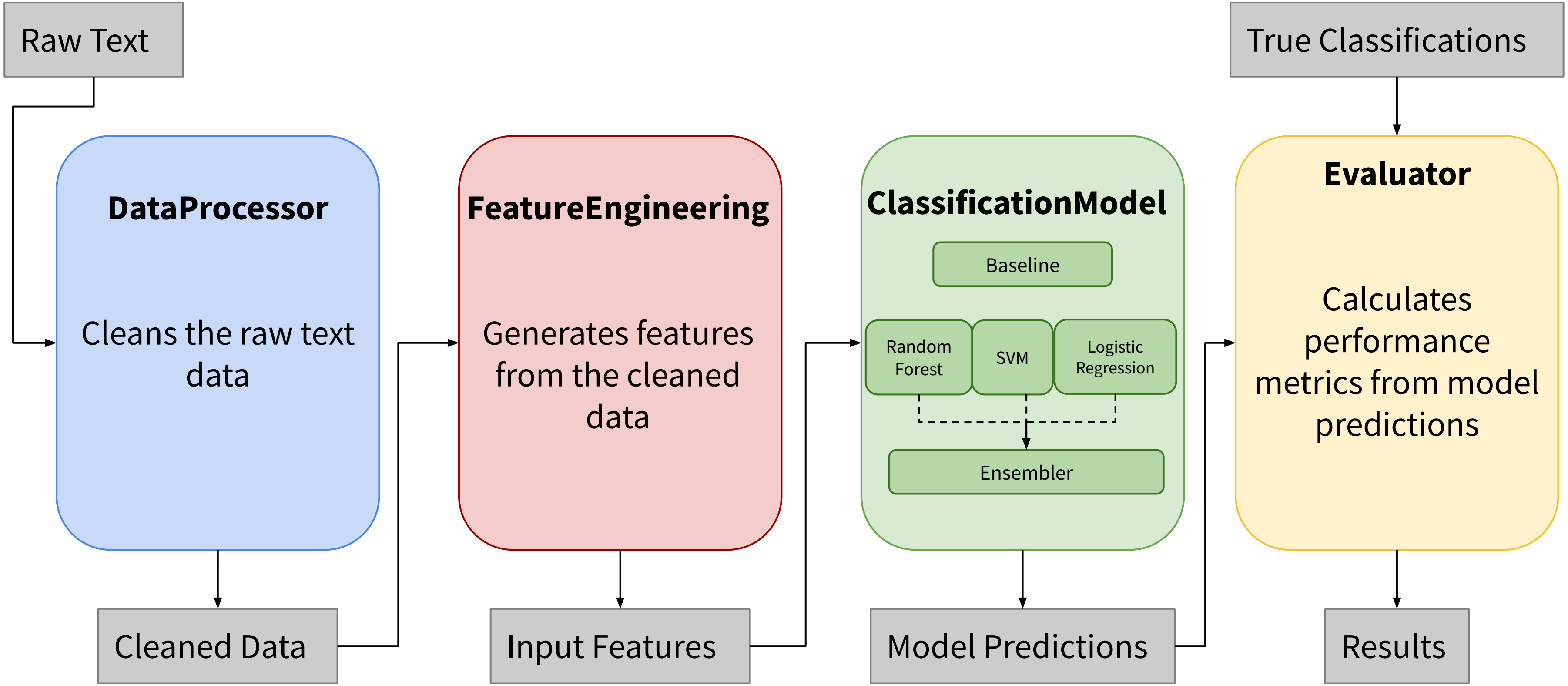
Adaptation Task

- Data is pre-labeled tweets in Spanish
- Binary classification (0/1) on Hate Speech, Aggression, and Target Group

What Changes?

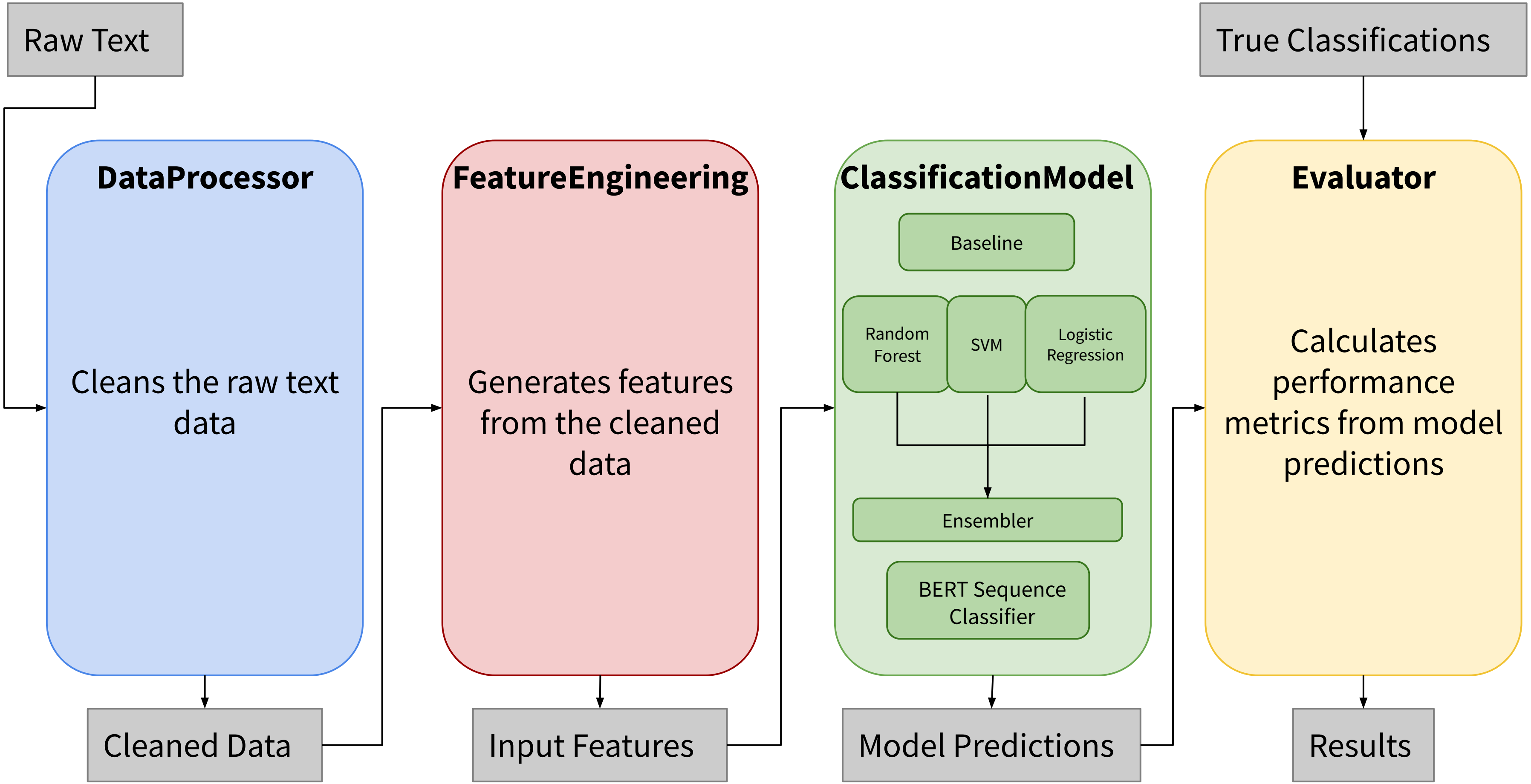
- Spellchecker must be able to handle Spanish and English
- Embedding features need different pretrained models
- Emotion Lexicon features need Spanish and English Lexicons
- Translator for a different approach to modeling Spanish

Our System



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

Our System


Raw Text
True Classifications
DataProcessor
Cleans the raw text data
FeatureEngineering
Generates features from the cleaned data
ClassificationModel
Baseline
Random Forest
SVM
Logistic Regression
Ensembler
BERT Sequence Classifier
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
- **Replaces typos in Spanish tweets with best guesses of correctly spelled word**

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

- (All Normalized): Punctuation Counts, NRC lexicon counts, Slang scores
- Embeddings:
    - Aggregated GloVe twitter embeddings
    - Google Universal Sentence Encoder embeddings
    - Aggregated BERTweet embeddings
- **NRC Lexicon extension– Use Logistic Regression & Glove embeddings to predict emotion labels for all words in dataset**
    - **Spanish and English**
- **Normalized NRC Lexicon counts using pre-built Spanish translation of NRC Lexicon**
- **Translator function to convert Spanish tweets into English for NRC Lexicon**
- **Spanish Emotional Sentiment Dictionary implemented to label words with sentiment strength**
- **Replace BERTweet embeddings with TwHIN-BERT to for multilingual use**
- **Implemented alternative GloVe embeddings if data in Spanish (trained on SBW corpus)**
- **Implemented Spanish Sentence Embeddings from sentence_transformers library**

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
NRCLex scores: [0.23,...]
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
- **Logistic Regression Classifier added**
  - **Predictions based on max number iterations and equal vs. balanced class weights**
- **Fine-tuned RoBERTa for Sequence Classification with raw data as input**
- **Ablation testing to ensure model tests on best feature combination***
- **Ensemble classifier uses either a Decision Tree or Logistic Regression Classifier to evaluate the results from best SVM, Random Forest, and Logistic Regression classifications in order to make an overall classification.**

INPUT

Percent capitalized: 0.029412
! count normalized: 0.385523
? count normalized: 0.428634
$ count normalized: 1.0
*** count normalized:** 0.46831
NRCLex scores: [0.23,...]
Slang score: 0
Universal sentence encoder embeddings: [0.045,...]
BERTweet embeddings: [-0.015,...]
Aggregate GloVe embeddings: [0.303,...]

INTERMEDIATE

Random Forest: HS+TR
SVM: HS
Logistic Regression: HS+TR

(These results are passed to a Logistic Regression or Decision Tree classifier)

OUTPUT

HS (hate speech): 1
TR (targeted): 1
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

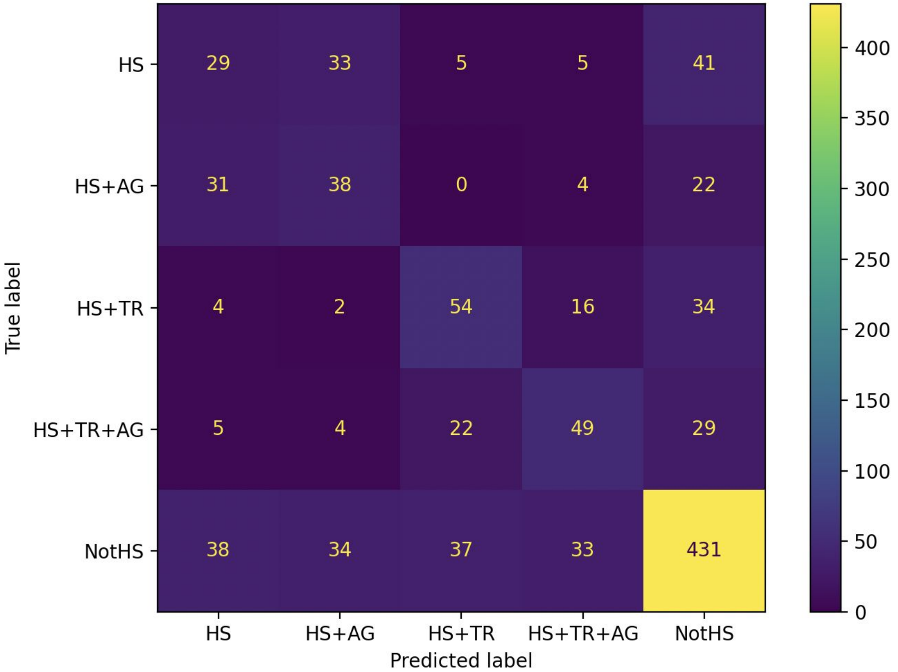
Evaluation and Results: All Models

Model	Data set	Macro F1	HS F1	TR F1	AG F1
SVM	English Validation	0.7088	0.7189	0.7621	0.6454
Random Forest	English Validation	0.6577	0.6580	0.7234	0.5915
Logistic Regression	English Validation	0.7022	0.7109	0.7463	0.6495
Ensemble-LR	English Validation	0.7148	0.7274	0.7655	0.6516
Ensemble-DT	English Validation	0.7030	0.7129	0.7421	0.6539
RoBERTa Sequence Classifier	English Validation	0.7404	0.7527	0.7852	0.6833
SVM	English Test	0.6982	0.7325	0.7301	0.6321
Random Forest	English Test	0.6389	0.6610	0.6486	0.6070
Logistic Regression	English Test	0.7006	0.7421	0.7226	0.6373
Ensemble-LR	English Test	0.7040	0.7482	0.7323	0.6315
Ensemble-DT	English Test	0.6843	0.7117	0.7090	0.6321

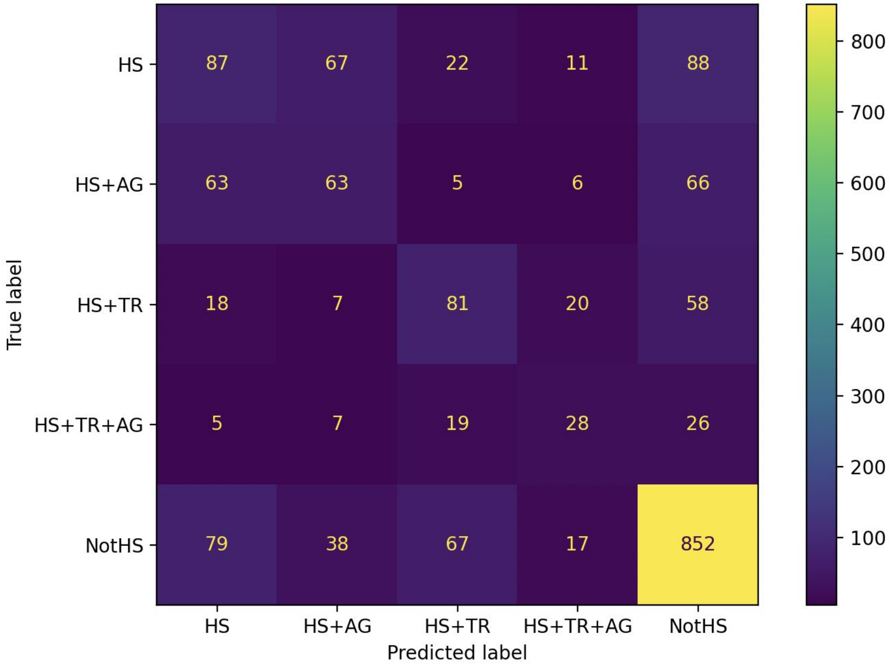
Model	Data set	Macro F1	HS F1	TR F1	AG F1
SVM	Spanish Validation	0.7369	0.7330	0.7748	0.7031
Random Forest	Spanish Validation	0.5724	0.4968	0.6689	0.5513
Logistic Regression	Spanish Validation	0.7585	0.7479	0.7929	0.7346
Ensemble-LR	Spanish Validation	0.7540	0.7476	0.7859	0.7284
Ensemble-DT	Spanish Validation	0.7496	0.7427	0.7839	0.7223
SVM	Spanish Test	0.7382	0.7168	0.7781	0.7198
Random Forest	Spanish Test	0.5638	0.4822	0.6360	0.5733
Logistic Regression	Spanish Test	0.7479	0.7430	0.7717	0.7290
Ensemble-LR	Spanish Test	0.7479	0.7390	0.7737	0.7311
Ensemble-DT	Spanish Test	0.7443	0.7349	0.7701	0.7279

Detailed Results for Top Performing English Model: Confusion Matrix

Validation Data

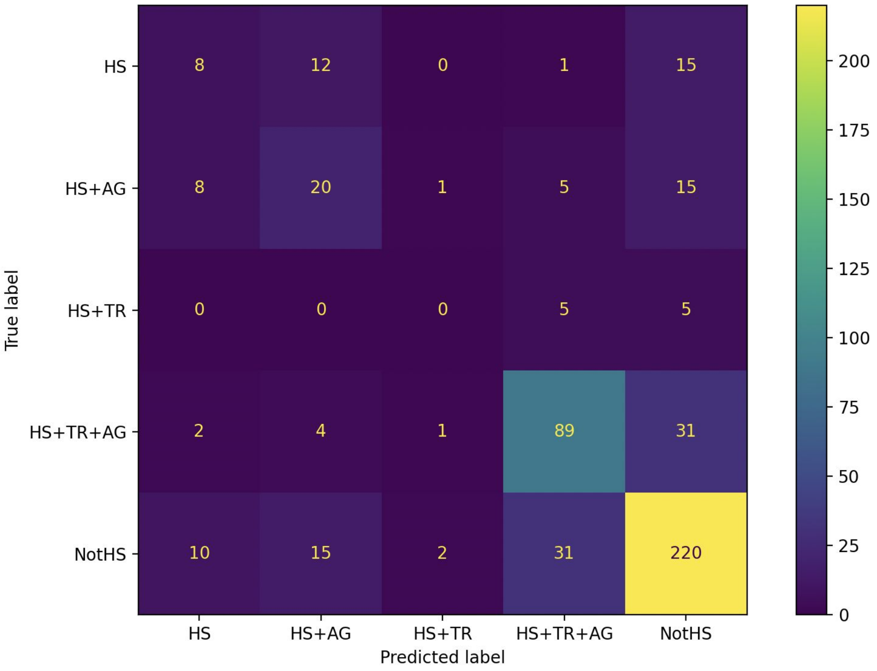


Test Data

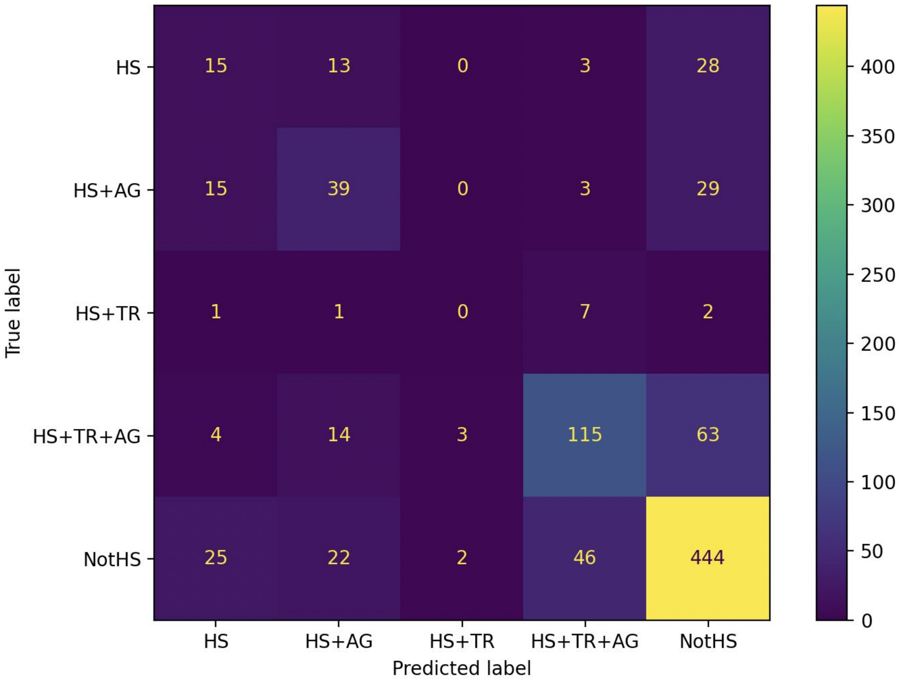


Detailed Results for Top Performing Spanish Model: Confusion Matrix

Validation Data



Test Data



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

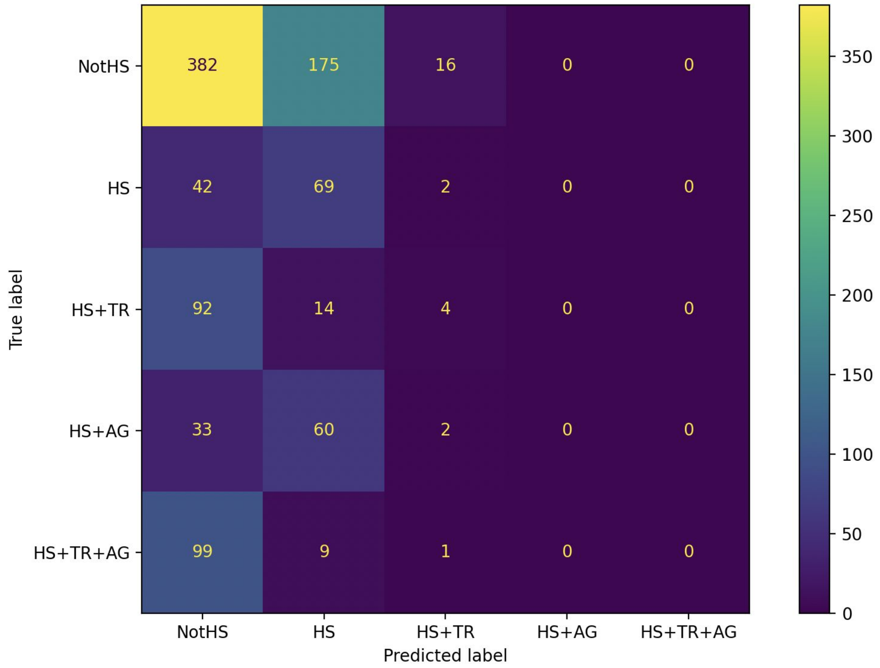
Detailed Results for Top Performing Model

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



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.