DEPARTMENT OF COMPUTER SCIENCE

Hats-off: A Multilingual Hate Speech Detector for English and Spanish Tweets

Yuexi (Tracy) Chen[1] Christine Herlihy[1] Rana Khalil[1] Jie Li[1] Yow-Ting Shiue[1] Chujun Song[1] Usama Younus[1]

University of Maryland, College Park, MD 20740 USA

Problem and Motivation

- **Problem:** Detection of hate speech on Twitter (against women and immigrants) in a multilingual perspective, for English and Spanish.
- **Motivation:**
  - ▷ Detect and address hateful content in un/weakly-moderated online settings (e.g., Twitter); use model output to design online or in-person interventions to protect targeted users from real-life aggressive actions.
  - ▷ Build models for lower-resource languages using majority language models as a foundation.

Objectives

The primary research question we seek to investigate in this work is: which combination of lexical and semantic features, along with model architecture, maximizes our ability to correctly classify whether or not a given tweet: (1) contains hate speech (against either immigrants or women) $\in \{0, 1\}$; (2) is aggressive $\in \{0, 1\}$; and (3) targets a generic group of people (assigned 0) or a specific individual (assigned 1). This research question has its roots in SemEval 2019's Task 5

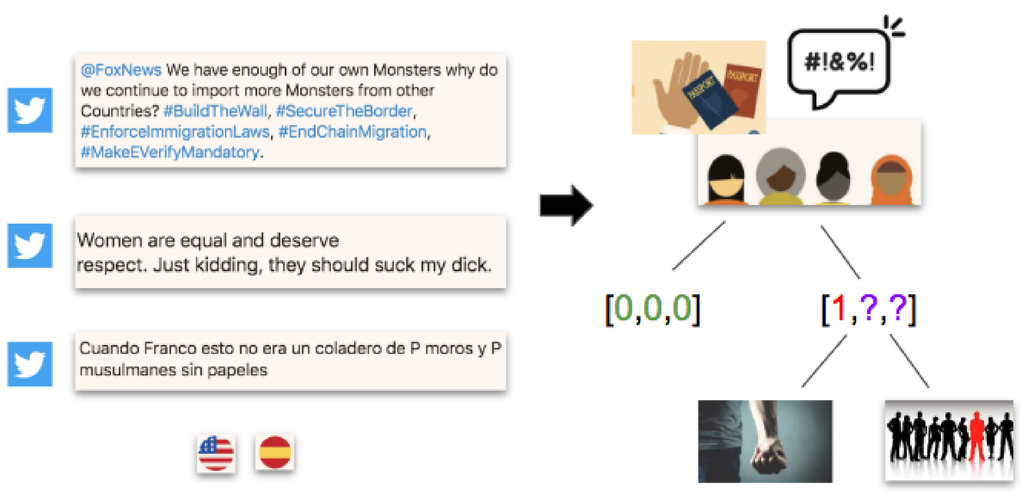


Figure: Our Classification Tasks

Data

- **Labeled Tweets**: SemEval 2019 Task 5 dataset [1]
  - ▷ **19, 600** labeled tweets (**13, 000** in English and **6, 600** in Spanish).
  - ▷ **9, 091** focus on immigrants; **10, 509** focus on women.
  - ▷ Language-specific train, validation, and test splits are provided.
- **Supplemental:** For feature extraction and engineering, we use language-specific lists of expletives and racial slurs.
- **Pretrained word embeddings:** We use GloVe+Common Crawl/GloVe+Twitter for English and fasttext+Wikipedia for Spanish.

Methodological Approach

- **Feature Engineering:** We extract lexical and syntactic features, including: ngrams; average embedding per tweet; Boolean and count values for expletives and slurs; and dependency relations.
- **Knowledge Augmentation:** We build directed graphs with edges weighted by frequency connecting tweet_id vertices from the training datasets to vertices representing mentions, hashtags, names, urls, emojis, slurs, and expletives. We then augment the text of each tweet by concatenating the text field(s) from all of its two-hop neighbors with edge weights $\geq$ a parameterized cutoff.
- **Non-Neural Models:** *baselines:* We implement three baselines: most frequent label classifier(MFC), linear SVM + TF-IDF, and a logistic regression classifier + n-gram. *An ensemble of classifiers:* We design a voting classifier, consisting of logistic regression, random forest and Naive Bayesian classifiers, each of which use real-valued, vectorized representations of tweets derived from pretrained word embeddings as features.
- **Neural Models:** We implement an LSTM network, using pre-trained `GloVe` embeddings for English, and `fastText` embeddings for Spanish. We evaluate performance for different combinations of hyperparameters: number of epochs, hidden size, embedding size, and activation function. We also explore and evaluate graph-augmented inputs.

System Architecture

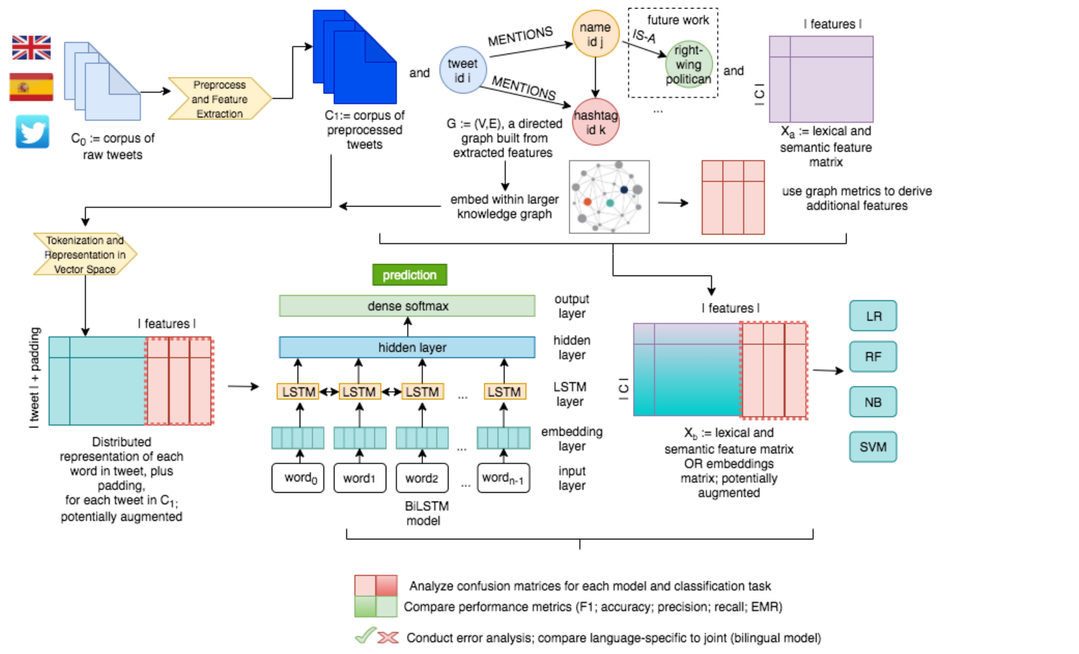


Figure: System Architecture

Results

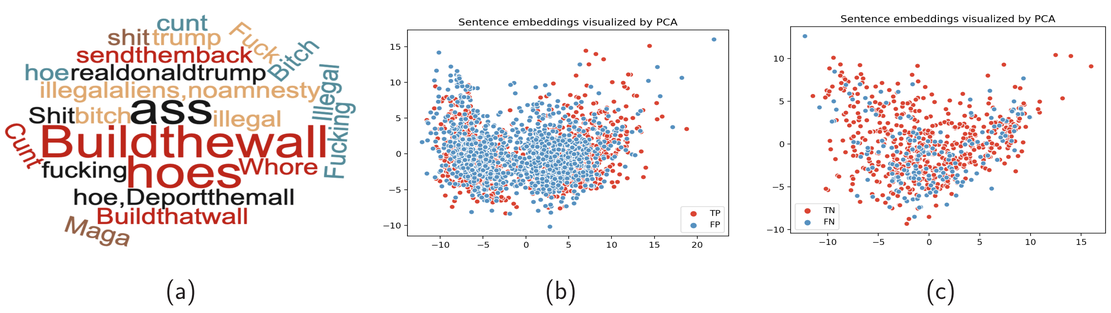
Language	English			Spanish		
Model	Baseline	Non-Neural	Neural	Baseline	Non-Neural	Neural
Task A	0.602	0.493	**0.626**	**0.708**	0.658	0.672
Task B	0.594	**0.605**	0.535	**0.725**	0.697	0.672

Table: Macro-F1 scores of official baselines, n-gram, our voting classifier and our neural models

For **English**, Task A, our LSTM model achieves the best macro-averaged F1 score, while the voting classifier outperforms in Task B. For **Spanish**, our ngram baseline model dominates both tasks. We hypothesize that our neural models under-performed due to the relative scarcity of training data, and also note that: (1) moving beyond the gains associated with ngrams requires deeper understanding of the context in which words are used than we are currently able to provide to our models; and (2) the ground-truth labels may be noisy, highlighting the difficulties of distinguishing between sarcasm, self-deprecation, and hate speech.

Error Analysis

To identify ngrams with predictive power, we quantitatively and qualitatively examine misclassified tweets. We compare top 10% frequent words in both true positive (TP) and false positive (FP) labelled tweets, and generate a wordcloud with the words that appear in the intersection of these sets. We also overlay principle components of GloVe embeddings of tweets with different labels in a 2D plot. While the wordclouds intuitively show us anchor words in prediction, the high overlap between TP and FP, TN and FN indicates the difficulty of separating tweets based solely on aggregated word embeddings.



(a) (b) (c)

Figure: a. Words from TP and FP Tweets. b. PCA of GloVe embeddings for TP/FP c. PCA of GloVe embeddings for TN/FN.

Future Work

- **Dataset Augmentation:** Ground our models and corpus-derived KG in a larger external knowledge base to encode additional semantic information.
- **Parameter Tuning:** Try joint embeddings; add attention.
- **Sensitivity Analysis:** Experiment with replacing anchor words.
- **Exploit Geographic Features:** Identify receiving-country anchor words (tied to sending country populations; borders/walls, etc.).

Sources Cited

[1] Basile, V. and et al. (2019, June). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th Intl. Workshop on Semantic Evaluation*, USA, pp. 54–63. Association for Computational Linguistics.