Sentiment Analysis of Code-Mixed Tweets

Shradha Agarwal (MT20123)
Tharun Suresh (MT20119)

Introduction

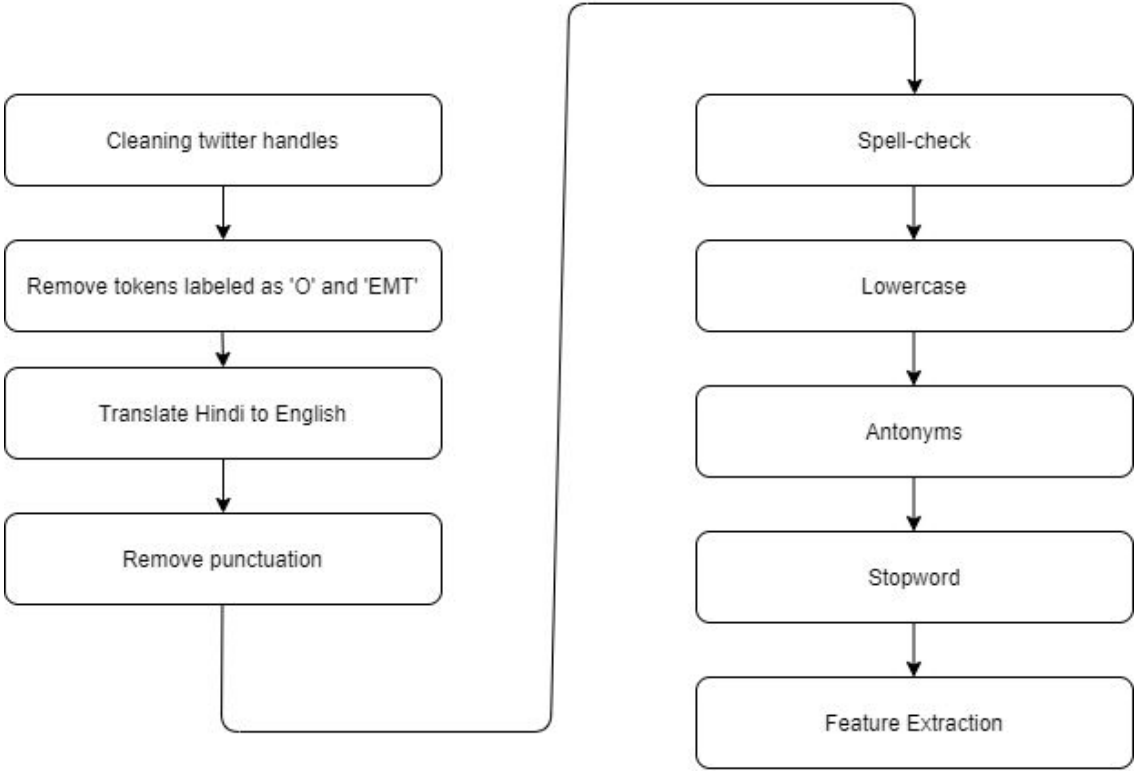
- SemEval-2020 Task 9 on Sentiment Analysis of CodeMixed Tweets (SentiMix 2020)
- Aim: To assign sentiment labels for Hinglish tweets.
- One of the important use case that can be tackled with sentiment analysis is hate-speech detection

- Code-mixing is one of the norms of multilingual communities where the users primarily transliterate to English language to express their own language.

- Since social media platforms are inherently multilingual environments and the presence of these platforms is now larger than ever, it is imperative challenging problem in the field of NLP
- The Hinglish (Hindi-English) corpora is annotated with word-level language identification and sentence-level sentiment labels- Positive, negative, neutral

- Dataset : Annotated Tweet samples of Hindi + English - 14k train, 3k validation, 3k test set

Pre-processing



Feature Extraction

- Word2Vec
  - Implemented various models.
  - Tuning for the optimal hyperparameters for different model implementations

- Created word embedding matrix for use along with Simple RNN
- Unique word list was used from the training set to set a baseline for the RNN

Model and Analysis

- Various models were implemented - Random Forest, SVM and Simple RNN
- Following parameters were tuned for Random Forest
  - Estimators - 500
- Following parameters for SVM:
  - C = 0.001, kernel = ‘rbf'
- Following was the best accuracy obtained from Simple RNN:
  - Parameters for RNN - 32 neuron units, Batchnormalization, dropout
  - Train accuracy - 70.8%
  - Validation accuracy - 65.3%
  - Test accuracy - 62.5%, F1 score - 63%
- In the SemEval-2020 Task 9, the baseline set for F1 score: 65.4%.
- Couldn’t reach the baseline of the competition.
- Reasons: 1) Error prone Google Translate API, 2) lack of sufficient feature extraction for native language words, 3) contextual coherence was not achieved across languages.

Contribution

- Shradha
  - Preprocessing steps, modifications
  - Random Forest and SVM
- Tharun
  - Feature extraction, modifications
  - Simple RNN