Detecting Stance in Tweets

Mani Kumar Reddy MT19065
Sarath Chandra Reddy MT19037
Vivek Reddy MT19038

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Problem Statement

- Automatic detection of stance has many applications including Information retrieval and there is active research going on in this domain of opinion mining.
- To develop automatic natural language systems that detect stance (favor or against) in tweets given the pre-target of interest.

Dataset Description

- We are provided 4,780 tweets with predefined targets of interest split into train and test datasets balanced by number of tweets across 3 different stances: favour, against & neither.
- Predefined targets of interest:
  - Atheism
  - Climate change concern
  - Feminist
  - Hillary Clinton
  - Legalization of Abortion
- Test file is provided with same structure as that of training dataset and model evaluation is done on that file to show how good the system is.

Preprocessing

- Stopword Removal
- Special Symbols removed with regular expressions and string attributes
- Lemmatization
- Stemming(for some models)
- Scale variations were normalized for Naive Bayes

Naive- Bayes

- Got the log priors and likelihood of the tweet being classified to one of the stances.
- Although laplace smoothing works better for Naive Bayes, we did chose it as hyper parameter (k-value) and calculated F1-score in k-space
- With the log priors and likelihoods across set of k-values for the training data, tested the tweet over the target of interest.
- Assigned the stance to tweet that leads to high probability.
- F1-score turns out to be : 50.2 (for k = 0.1)

Logistic Regression

- Used TF -IDF word embeddings for the tweet data after its preprocessing.
- To get best optimal values applied Grid Search Cv with validation set.

  C = 11.28

  Penalty = l2 regularization

- F1 score for train data : 96.2
- F1 score for test data : 54.8

SVM

- After preprocessing of the tweets, got the TF-IDF representation for every tweet.
- The vectorized tweet obtained was trained on SVM model for stance detection.
- GridSearchCV was applied to look for the best params of the model.

alpha = 1e-10 (used elastic-net regularization with L1 ratio as 0.6)

F1-Score: 56 (avg. of Macro Avgs)

LSTM

- Each word in the sentences is replaced by the rank of word in the vocabulary.
- We have also done padding which makes length of all sentences to be equal.
- The rank of unseen word in the test set is assigned a particular number or is ignored.
- Maximum length of the sentence considered to be 30 words (after padding).
- **OPTIMAL PARAMETERS:**
  - LSTM units :12.
  - Dropout rate : 0.8
  - Activation in the output layer : Softmax
- **Optimal F1-Score obtained : 74.66**

F1-Scores for different models

- Naive Bayes : 50.2
- Logistic Regression : 54.8
- Support Vector Machines : 56
- LSTM : 74.66

Conclusion

Approached different ML models and found LSTM to beat the F1-scores of the rest models because the dependencies are being preserved in the LSTM where the later approaches used vectorized operation on tweet for the model to be trained.