Updated

CSE543 - Machine Learning

Sentiment Analysis to Classify Online Abuse

Motivation

With the increase in online presence, there is an equal increase in responsibility to moderate these kinds of activities online, so as to provide a safer environment for everyone. Automated filters to detect such activities have repeatedly failed to detect these colluded users, which motivates us to develop a learning model which can detect the negativity and classify them. In our model, we find negative comments or intent and try to classify them or group them with other similar comments.

Dataset

We are going to use a mixture of datasets as we faced a huge problem in the datasets where it says that few datasets are not labelled.

1. Data generated for the Wikipedia Detox project is available under free licenses on the Wikipedia Talk Corpus
2. Hate Speech Twitter annotation: http://research.cs.wisc.edu/bullying/data/bullyingV3.0.zip
3. https://github.com/zeerakw/hatespeech

Preprocessing

As we are dealing with textual data, there is a lot of preprocessing required

- Set all characters to lowercase
- Remove numeric characters
- Remove punctuation
- Remove stop words
- Tokenisation technique
- Lemmatization technique
- Noise removal technique
- Stemming technique

Learning Techniques

- *Baseline*
  - Naive Bayes, Logistic Regression
- *Advanced*
  - MultiLayer-Perceptron (MLP)
  - Recurrent Neural Network (LSTM)
  - Convolutional Neural Network (CNN)

Strategy for Model Selection

First we will implement using the Logistic Regression and Naive Bayes models to set our baseline after which we will try to build a better model to get a higher accuracy rate by trying to implement the advanced learning techniques mentioned above.

Tuning Hyperparameters

K-Fold Cross-Validation

Training Approaches

Gradient Ascent or Descent for Logistic Regression

Classifications *(Tentative)*

We will mainly classify if it is an abuse and then we will also try to differentiate the type of abuse depending on the dataset.

- Personal Attack
- Toxicity
- Accusation
- Cyberbullying

Performance Improvement

Performance improvement process is more empirical than theoretical. We are going to try out different approaches which can improve our performance and analysing on why we have achieved such an accuracy. Few of the approaches we plan to implement:

- ReLu
- TF-IDF
- n-grams
- Speaker's Native-Language consideration

Testing Approach

Few additional test dataset will be extracted from Twitter for different geological regions and each sample will be classified by a minimum of 3 members to detect the label and the maximum votes for a label will be noted as the label. The trained model on the initial dataset is tested on the Demographic datasets. And we will try to improve the performance by trying different enhancement techniques.

Evaluation Metrics

Confusion Matrix (Error Matrix): We use confusion matrix to give us a gist of predictions in a layout manner using actual values and predicted values on columns and rows respectively. This helps us understand if the model is confusing among two classes.

Receiver operating characteristic (ROC): We use ROC to plot using the true-positive rate with false-positive rates with different thresholds to evaluate the performance of the model.

Deliverables

All the team members will be contributing equally in understanding the problem statement and in finding the feasible solutions. There are three advanced models mentioned above, each of which will be implemented individually.

William Scott - MT18026
K.Srivatsava - MT18054
Subhani Shaik - MT18117