IT412: Natural Language Processing

Assignment 9: Text Classification

Instructor: Prasenjit Majumder

Learning Outcome: At the end of this assignment you will learn to create text classifier and their evaluation.

1 Text Classification

Classification is a supervised machine learning problem. Classification can be binary i.e given mail is spam or not, multi-class i.e categorize a given news as sports, entertainment, politics etc. Aim of **Text Classification** is to automatically classify a given document. Text classification is the task of assigning a set of predefined categories to free-text. Text classifiers can be used to organize, structure, and categorize pretty much anything. Few examples of text classification are span detection, sentiment analysis etc. Figure 1 shows end-to-end text classification Pipeline

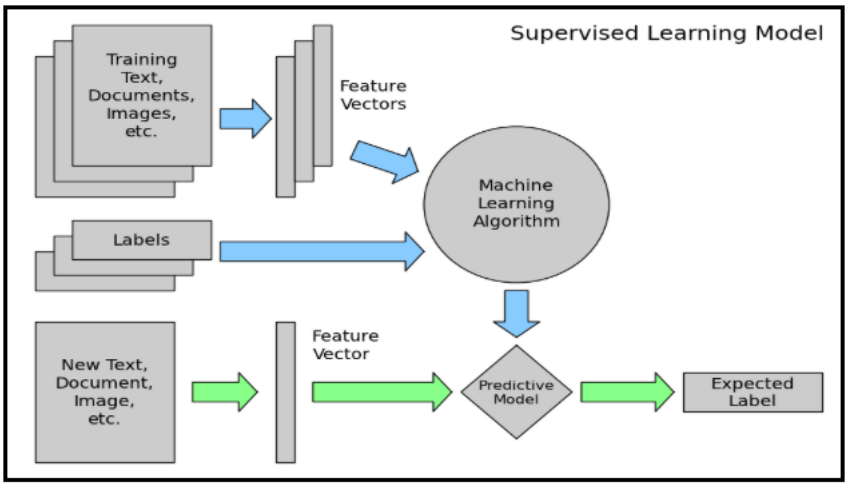

Figure 1: Text Classification Pipeline[1]

2 Implementation

2.1 Dataset

- You are provided with *Hate Speech and Offensive Content Identification in Indo-European Languages* (HASOC)[2] dataset. Competition had three sub-tasks as shown in Figure 2.

2.2 Exercise

1. Perform text preprocessing for given data as required.
2. Create classifier whose features will be represented using TF-IDF for all three sub-tasks. Calculate MacroF1 score for test data on trained model.
3. Create classifier whose features will be represented using any Pre-trained Word vectors i.e Word2vec,GloVe and FastText for all three sub-tasks. Calculate MacroF1 score for test data on trained model.

[1] https://medium.com/jatana/report-on-text-classification-using-cnn-rnn-han-f0e887214d5f
[2] https://hasocfire.github.io/hasoc/2019/index.html

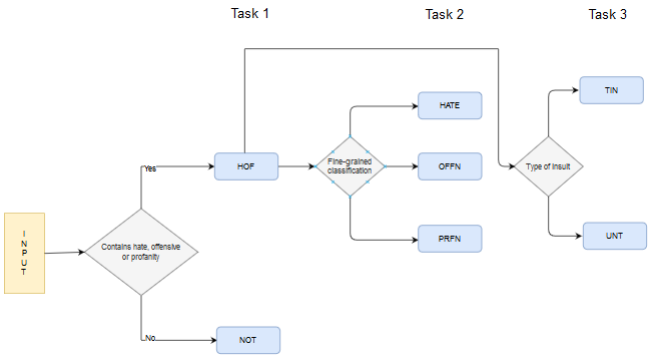

Figure 2: HASOC Challenge

Note:You can use any classifier i.e Logistic Regression, SVM, Random Forest etc. Pre-trained Embedings can be of any dimension.

3 References

1. HASOC Working Notes
2. TF-IDF