IRE Project

First Deliverable - Scope

Humour detection in Yelp reviews using Deep Learning

Ankita Jain (201505536), Anurag Gupta (201301109), Shraddhan Jain (201330174)

Problem Statement -

Our problem is to predict community-determined humor in Yelp reviews. Yelp reviews can be voted as funny and, given a review, it is our task to label it as funny or not, using machine learning techniques.

Applications -

Yelp is a service whose utility relies on being able to provide an accurate, crowd-sourced opinion of a place of business. So it's important to determine which reviews provide valuable insight into a business. Humor could play an important role in this determination and help consumers to get to informative reviews and not waste time on reviews which have nothing to provide. So, our project would help in correctly identifying humor.

Challenges -

We cannot have an idea whether or not our models can be extended out-of-domain. This is because there is no labeled corpus of funny texts – partially due to the subjectivity associated with imposing a binary outcome on something such as humor. Some work has been done applying similar methods to Twitter, but no large, labelled datasets exist for thorough validation.

Second Deliverables Details -

Pre-processing -

1. Tokenize the reviews, without removing stop words or punctuations, as they are quite expressive in the context of Yelp.
2. Build word vectors using gensim package.

Shallow Algorithms -

Explore shallow algorithms before moving to Deep Learning. We will apply SVM on bag of words.

Evaluation -

We convert the above mentioned problem into the binary classification regime. After balancing the classes according to funny and not-funny, we optimize the classical log-likelihood and report the classification accuracy.

Third Deliverables Details -

Implement the following deep learning techniques -

1. Deep Feedforward Networks on two types of architectures – a classic, deep rectifier network with heavy dropout, and a slightly more modern, shallower MaxOut.
2. Convolutional Neural Networks

Tools -

1. Gensim Package - Build word vectors
2. NLTK - Tokenization
3. Keras library - Deep Learning Techniques
4. Scikit library - SVM Classifier

References -

1. Oliveira and Rodrigo, Humor Detection in Yelp reviews
   URL: https://cs224d.stanford.edu/reports/OliveiraLuke.pdf