**CS 491: Introduction to Machine Learning Spring 2014**

**Project Proposal**

**Team Members:**

* Abhinaya Balasubramanian
* Balachandar Venkataramani
* Sruti Bhagavatula

**Problem Statement:**

We aim to classify the sentiment expressed by the movie reviews obtained from Rotten Tomatoes. ‘Rotten tomatoes’ is a film review aggregator where users provide the review for the movies. We intend to build a classification model using Machine Learning techniques that will classify a review into positive, negative or neutral class.

**Dataset Description:**

This problem is a part of ongoing Machine Learning competition from Kaggle. They have also provided the labelled datasets. The datasets originally consists of 5 classes (Positive, Slightly Positive, Neutral, Slightly Negative, and Negative). We have grouped Slightly Positive and Positive into a single ‘Positive’ class and also grouped ‘Slightly Negative’ and ‘Negative’ into a single ‘Negative’ class thereby reducing the complexity due to time constraints. Each review is split into sentences and each sentence in turn is split into several phrases. Each of these phrases has a sentiment label. There are 8544 sentences and 156060 phrases in the training set.

**Machine Learning Techniques:**

**Preprocessing:**

We will use regular preprocessing techniques like stopword removal, stemming, removal of long repeating characters, etc.

**NLP techniques:**

* We also intend to use some of the NLP tools like Stanford Parser to extract adjectives and adverbs.
* Also, tools like SentiWordNet to extract words that express sentiments (opinions).
* Tracking direction changes in the texts (e.g., ‘although’,’however’,etc)

**Other useful techniques:**

* Some of the other useful approaches like Bigram models with the help of Bigram Collocation finder in Python NLTK kit which helps in getting top ‘n’ bigrams selected based upon chi-square test.
* Representing the texts as a Bag of Words model using a Binary scheme(tracking only the presence or absence of a word)

**Feature Selection:**

We use effective feature selection techniques for removing irrelevant features that do not contribute to the effective classification and to avoid overfitting. Some of the techniques that we intend to use are:

* Chi-Square Test
* Information Gain

We may just use one of them or the combination of the results obtained from the above two.

**Machine Learning Algorithms:**

We intend to use Supervised Learning algorithms:

* Support Vector Machine [[1]](#footnote-2)
* Naïve Bayes – Bayesian Multinet Classifier [[2]](#footnote-3)
* Random Forest

We then compare the results and use the algorithm which best fits our model.

**Evaluation:**

We evaluate the model using the following measures:

* Precision
* Recall
* Accuracy

1. <http://www.infoautoclassification.org/public/articles/Ikonomakis-et.-al._Text-Classification-Using-Machine-Learning-Techniques.pdf> [↑](#footnote-ref-2)
2. <http://www.infoautoclassification.org/public/articles/Ikonomakis-et.-al._Text-Classification-Using-Machine-Learning-Techniques.pdf> [↑](#footnote-ref-3)