Automatically Selecting Best Features During Semantic Similarity Measure

**Lasang Jimba Tamang**

University of Memphis, Memphis, Tennessee

ljtamang@memphis.edu

**Abstract.** Feature selection during semantic similarity measure is one of the most important task. The performance of model depends on both the selection of features and appropriate number of features. In this project, we select best Fi number of features for i <= total number of features using chi square feature selection, build model, and evaluate the model, and finally the feature Fi that yields best performance.

Log and Source code

Progress log can be found [here](https://livememphis-my.sharepoint.com/personal/ljtamang_memphis_edu/_layouts/15/guestaccess.aspx?guestaccesstoken=PgGturx1D3%2fxK5e17F3%2fIImhOqWNyHRfNGNy0AGhtw8%3d&docid=2_114b47eafa07540ee88a75ba489b1dd8c&rev=1) and source code can be viewed [here](https://github.com/ljtamang/nlp_proj) .

1 Introduction

Semantic similarity measure uses different classification and regression machine learning algorithm. The performance of these algorithm depends largely on the feature set we use. However, choosing the feature set that best contribute to model performance is tedious job.

In this paper, we automatically select the best features that gives the best performance.

2 Approach

In our approach, we iteratively select different number of best features using chi-squared test, build regression model and classification model using those features set, and evaluate the performance using 10-fold cross validation. The feature corresponding to best average performance of regression and classification model performance during evaluation is reported as best features to be used for building model.

2.1 Data Set

Our data set consists of 4514 row with 51 columns. The first 50 column consists of score for different kind of features and the last column consists of similarity score between two sentence. We also call the last column as gold value. This data is generated by **Deeptutor Lab**, Institute of Intelligent System at University of Memphis, during SemEval-2017 International workshop on semantic evaluation. We call this **dataset 1.**

We take dataset 1, converting its continuous gold score value into discrete value and replace all negative values to 0. This conversion is essential since chi-squared test can be performed only on discrete gold value and data with non-negative values. We call this **dataset 2.**

2.2 Feature Set Formation

We use dataset 2 and perform feature selection for i<= 50 times. At each i iteration, we ask the chi-square method to output i best features and called it Fi. In this way, we will have 50 different feature set where each Fi feature set consists of i best features.

2.3 Model Formation

We take 50 different features set from step 2.2 during the model formation and for each feature set Fi where i<=50, we form both classification and regression model CMi and RMi respectively. Classification model target value is set from from dataset 2 and regression model target value is set from data set 1. However, both model is trained using data set 1.

2.4 Best Features Selection

Each CMi and RMi model is evaluated using 10-fold cross validation. The feature set Fi is selected as best feature that yield best average accuracy of CMi and RMI.