Classifying responses on online discussion forums

1.Introduction About Problem :

There are online discussion forums covering an endless variety of topics . Each forum has many topic-specific questions with expert answers. However, the best answer is often buried somewhere in the discussion thread.Given a question on online forum ,we focus on the problem of classifying each response to the question as an answer or a non-answer. where the answer label indicates if the response made a reasonable attempt at answering the question(even if the answer is incorrect).

2.Approach :

I trained a Decision tree that labels responses to forum questions as answers or non-answers.The answer label indicates any response that attempts to answer the question.

Our Decision Tree model largely based on nested if-else conditions.Given a training set it would make branches such that maximum information is gained at each branch and there by reducing total entropy at each branch.This branching continues till leaf node reaches pure node or Information gain is zero or specified condition.

Model :

Model will make branches in such a way that maximum information gain at each branch.For every brach it makes there should be decrease in entropy.To caluculate Information gain we used Gini Impurity.

Mathematical Formula Bearing Gini Impurity:

$$I_G(Y)=1-\Sigma^{n}_{i=1}(p(y_i))^2$$

Where i={1,2}

I_G=Gini Impurity

$Y_i=\{0,1\}$

Previous work Reveals that if we won't give right depth the model may overfit for entire data.However,it may fail for test set.By tuning the depth parameter we can find right depth for the Model.

Experiment :

I trained Decision Tree classifier using responses from online forum.

Training Data :

Our training data consists of responses from forum threads.As a simple example consider the responses"The sky is blue","I am not sure","why don't you just google it ?","I think its Purple".These statements lead to Three training inputs.

The first,fourth are answers while second and third responses are non-answers.

Note that last response is incorrect answer ,but it still constitutes an answer.

Data Acquisition and Preprocessing :

I built a web crawler using selenium and java and downloaded 24174 question-response pairs from one online forum.we hand labelled data as answer or non answer.Myself and Project mate hand labelled each response.I randomized the data and then split into train (63%),cv(16%),and test(20%).

	% of total dataset	No.of.Q&A pairs
Train	63	15729
Cv	16	3933
Test	20	4916

I preprocessed the question-response pairs by applying the following normalizations :

- Remove Urls.
- Remove HTML tags.
- Replacing Emoji and Emoticons
- Clean Contractions.
- Chat Word conversion.
- Replace repetition punctuation.
- Convert Special characters .
- Remove Punctuation.

- Replace Negations.
- Remove Numbers
- Convert Text to lower case
- Remove Stop words
- Lemmatize Words
- Correct Spelling

Evaluation Metric :

I used the number of correctly classified question response pair from the test set as my evaluation metric.

Conclusion :

Our Decision Tree classifier Really performed well with 91.3% accuracy on test set.It has Recall of 94 % and Precision of 88%.we can improve accuracy even more with Bagging approach with different Base models.