KweriME
A Q&A based model which predicts the accepted answers of questions in CQA sites

Kaushik N
Vasundhara Singh
Shalini Shekar
Harini

Prof. Evlin VidyuLatha P
Batch No. - 16

April 13, 2018

Problem Statement / Definition

- The dissertation aims to address the issues that reside in the community based Q&A websites with KweriME, a reputation based QA system which employs a category and theme based reputation management system to evaluate users willingness and capability to answer various kinds of questions, while at the same time improving the response latency and answer quality.

Motivation of the Work

Motivation for our work includes the following:

- Seeking information on the internet has become a daily part of our lives. This inspired us to build an efficient system which reduces delay as well as gives us additional information on the content.Most of the Q&A systems experience a delay which we try to solve by forwarding the questions to experts thus, ensuring improved answer quality and also incorporate an anti-spammer control to filter out irrelevant and advertising verbiage to optimize user interaction.Instead of going through multiple questions, the asker will be satisfied with the best and optimal answer, thus, saving time and energy of the asker.
- Previous works include the asker to review the answers manually, thus consuming a lot of time as the answers being posted on these sites grow rapidly. This becomes a tedious job which can be solved by an efficient system. Exploring the talent & knowledge of the answerer by labeling them as experts & providing them a good platform to showcase their talent.

Literature Survey

Topical interest and Recommending the best answer are the talk of the town which has attracted many researchers interest.

- In [1], the activeness of users has been explored in CQA.They have shown how badges and reputation scores are related to the activeness in different forums based on statistical analysis.Yuhua Lin et al. discussed on clustering the users of Stack Overflow into four clusters namely naive,surpassing, experts and out shiners based on characteristics accounting various metrics by using machine learning algorithm in order to predict the users activities.
- In [2], Tirath Prasad Sahu et al. has worked on the goal of uncovering topic interest, main discussion topics and technology trends over time with the help of statistical topic modeling and also they have worked on the goal of uncovering topic interest, main discussion topics and technology trends over time with the help of statistical topic modeling. In [2], the Questions posted on Stack Overflow has been analyzed both quantitatively and qualitatively in order to improve the success of CQA.

Methodology

The approach towards the problem definition is divided into following major components:

- **Data Selection :**

  We perform an extensive empirical analysis on a QA based online community dataset to answer the three research questions. Our findings will suggest that:
  i) Prior involvement of the answerer on question tags and topics increases the chance to give the answer for that question.
  ii) Expertise will increase the chance in acceptance of the answer.
  iii) Topical compatibility between the question and answer increases the satisfaction of asker or community with that answer.
  Furthermore, we use various statistical methods in order to implement the algorithms to predict acceptability of the answer by the asker or community.

Methodology

For instance, we consider the placeholder for the aforementioned online community to be StackOverflow, which is an interactive CQA site for exchanging the knowledge in the software engineering field. It provides a wide variety of functionality for users to gain knowledge in their respective domains. In StackOverflow, there are many questions from various topics related to programming. There are about 10M questions, 17M answers, 4.5M users and 42K tags in the StackOverflow till May 31, 2015. StackOverflow offers its data publicly which is available through Stack Exchange Data Explorer and XML format data dump under creative common licence. The statistics about the dataset relevant to our study are to be analyzed and studied in order to make further inferences on prediction accuracy.

Methodology

- **Applying Machine Learning techniques :**

  Numerous number of Machine Learning techniques are available which can be used for processing of data .We model a prediction system as a binary classifier, which classifies an answer as an accepted answer or not an accepted answer. We next use classification algorithms based on Bayes rule such as Naive Bayes to predict acceptability of the answer by the asker or community and for the task of binary classification, in which the generative model is utilized for modelling question and answer based on Gaussian distribution. The evaluation parameter of the classifier reveals that the results are remarkable in predicting the answer acceptability.

- **Prediction :**

  Using Machine Learning we predict accepted answers for any post.

Detailed Design

- **STEP 1: DATA COLLECTION**

We used Stack Exchange Data Explorer for running arbitrary queries against public data from the Stack Exchange network and also gathered data dumps of random timeframe from StackExchange archives.
The dataset consists of posts from Stackoverflow website which can be obtained in .csv format from StackExchange archives
We have merged, transformed , preprocessed our data to fill in missing values and to deal with categorical variables.
At present, Stackoverflow consists of over 15M questions and 24M answers with over 64M comments.
Owing to the huge dataset to deal with, We decided to take a subset of these posts around 60,000 questions which were tagged as python or python-2.x or python-3.x.

- **STEP 2: DATA VISUALIZATION**

Exploratory Data Analysis (EDA) is an approach to analyze the data set with respect to their characteristics and present it in the form of visual methods through the use of various kinds of graphs.

The libraries that we used for EDA implementation are pandas and seaborn.

- Pandas is an open source library which provides tools for data analysis in the Python programming language.
- Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.

- **STEP 3: FEATURE ENGINEERING**

It is performed based on extensive literature survey in terms of research questions.

We formulate the following four research questions :

- **Who provide answers to the question on stack overflow**
- **How much expertise does the answerer has in :**
  (i) same domain (ii) different domain?
- **To what extent the answer is relevant to the question?**
- **How the best answer is selected amongst a set of answers**

1. Who provide answers to the question on stack overflow

- In StackOverflow, there are many questions from various topics related to programming.
- There are about 10M questions, 17M answers, 4.5M users and 42K tags in the StackOverflow till May 31, 2015.
- We try to identify answerer based on the metrics such as activeness and the domain knowledge of a particular user in a specific domain

2. How much expertise does the answerer has in (i) Same domain (ii) Different domain of the question?

The knowledge of the answerer can be derived from the metrics such as number of up-votes,reputation and percentage of accepted answer.

As the knowledge of the answerer increases,the expertise level also increases, consequently the respective answers are likely to be accepted.

- **ANSWERER SCORE**
- **REPUTATION**
- **ACCEPTED ANSWER**

3. To what extent the answer is relevant to the question asked?

We find the compatibility of the answer with the question using various topics to meet the satisfaction level of the asker

We calculate the relevancy of the answer given by each answerer to the question asked:

- **TOPIC RELEVANCY or SIMILARITY**

4. How the BEST answer is accepted amongst a set of answers?

We compile the answer of above three research questions, which acts as the baseline in order to answer this research question. Based on a posts feature vector that we have extracted, the answer is classified as accepted or not. We plan to employ NaiveBayes for the task of binary classification

Apart from this we try to implement other contextual features such as :

- **TIME SPAN OF ANSWER**
- **ANSWER SCORE**
- **SENTIMENT ANALYSIS**
  (for comments in the measure of polarity)
- **READABILITY**

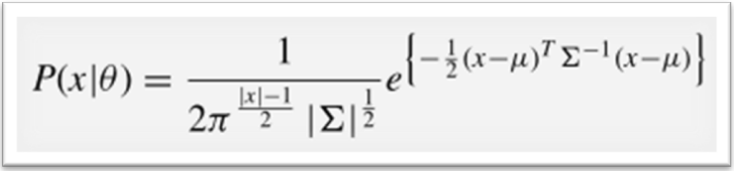
Classification Models

We model a prediction system as a binary classifier, which classifies an answer as an accepted answer or not an accepted answer.

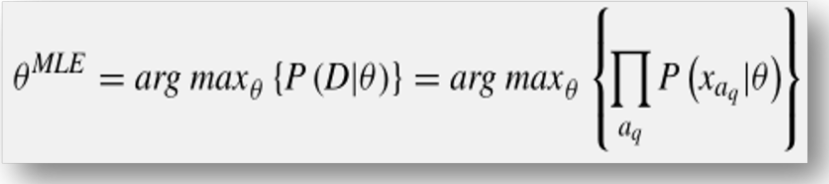
- **Gaussian Naive Bayes**
- **Decision trees**
- **Dummy classifier**

Guassian Naive Bayes

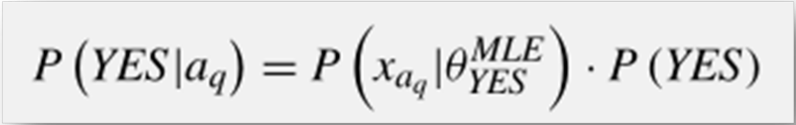
The model assumes that answers are Gaussian distributed in terms of their acceptance.



$$P(x|\theta) = \frac{1}{2\pi^{\frac{|x|-1}{2}} |\Sigma|^{\frac{1}{2}}} e^{\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}}$$



$$\theta^{MLE} = arg\ max_{\theta} \{P(D|\theta)\} = arg\ max_{\theta} \left\{ \prod_{a_q} P\left(x_{a_q}|\theta\right) \right\}$$



$$P(YES|a_q) = P\left(x_{a_q}|\theta^{MLE}_{YES}\right) \cdot P(YES)$$

Decision tree

A decision tree is a set of rules used to classify data into categories. The key idea is that the procedure to create decision trees is recursive.
For a set (S) of observations, the following algorithm is applied:

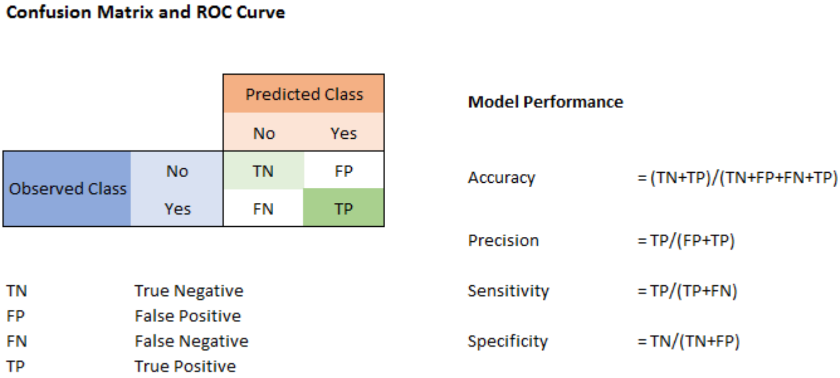
- **If every observation in S is the same class or if S is very small, the tree becomes an endpoint, labeled with the most frequent class.**
- **If S is too large and it contains more than one class, find the best* rule based on one feature (e.g., "isweight $>$ 150?") to split it into subsets, one for each class.**

Performance evaluation metrics

These evaluation measures are generally used to assess the performance of the classification where the dataset is imbalanced as we have.

- **Accuracy**
- **Precision**
- **Recall**
- **F1 score**
- **ROC-AUC curve**

Confusion matrix



Confusion Matrix and ROC Curve

		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

Model Performance

Accuracy	= (TN+TP)/(TN+FP+FN+TP)
Precision	= TP/(FP+TP)
Sensitivity	= TP/(TP+FN)
Specificity	= TN/(TN+FP)

F1 score

F1 Score: This is a weighted average of the true positive rate (recall) and precision.
We consider F1 score as our measure of test accuracy, which considers precision and recall to compute the score.

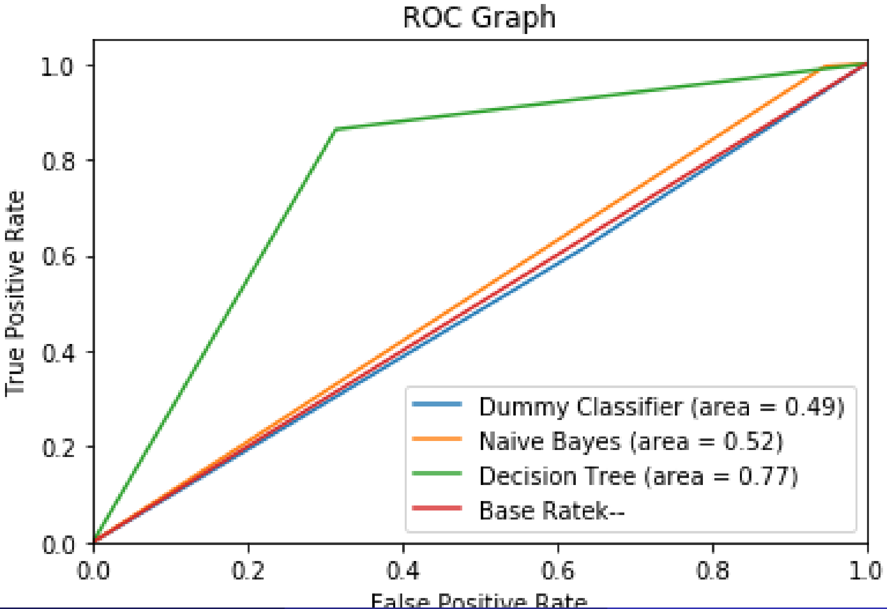
$$F_1 = 2 \cdot \frac{1}{\frac{1}{\mathrm{recall}} + \frac{1}{\mathrm{precision}}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

ROC Curve

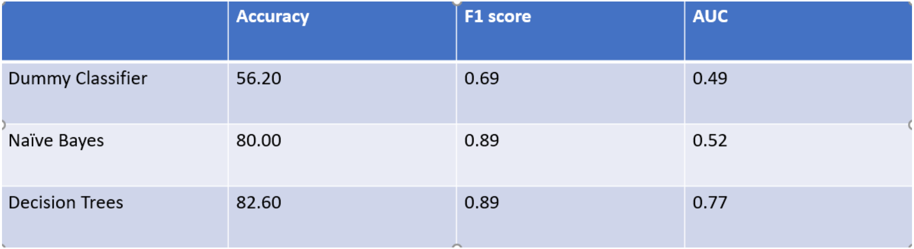
This is a commonly used graph that summarizes the performance of a classifier over all possible thresholds.

AUC: This area equals the probability that a randomly chosen positive example ranks above a randomly chosen negative example.

ROC Curve

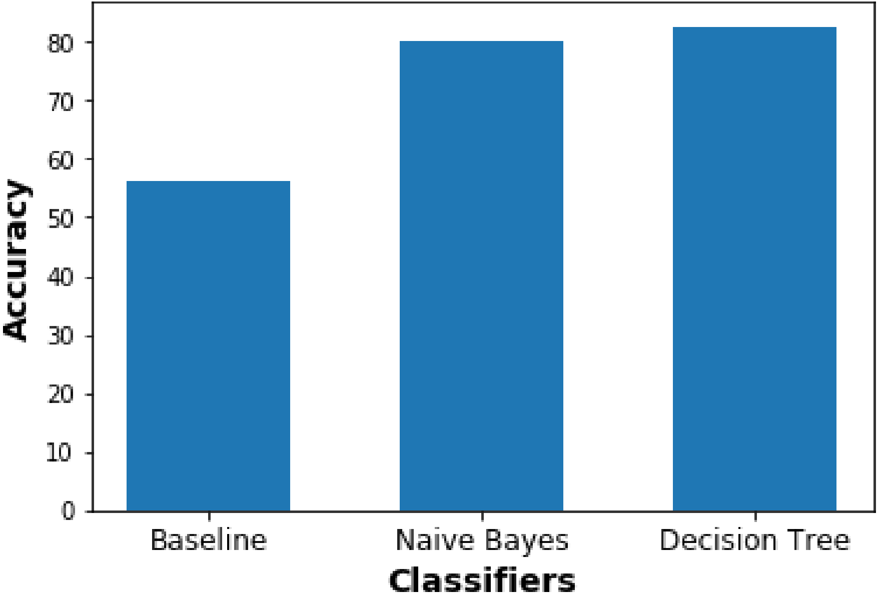

ROC Graph
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Dummy Classifier (area = 0.49)
Naive Bayes (area = 0.52)
Decision Tree (area = 0.77)
Base Ratek--

Results

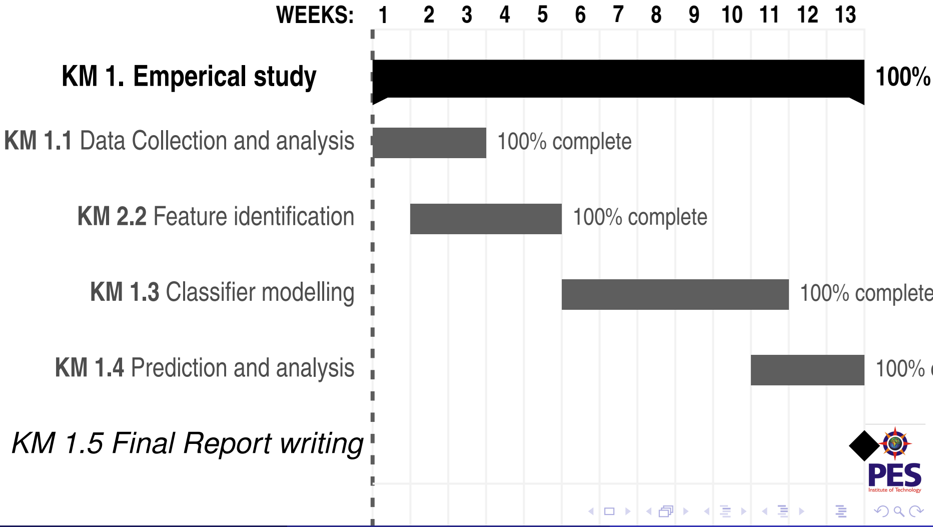


	Accuracy	F1 score	AUC
Dummy Classifier	56.20	0.69	0.49
Naïve Bayes	80.00	0.89	0.52
Decision Trees	82.60	0.89	0.77

Results


Accuracy
80
70
60
50
40
30
20
10
0
Baseline
Naive Bayes
Decision Tree
Classifiers

Time line of completion of project from Nov 2017-April 10 2018(Gantt Charts).



Future scope

We study and analyze the answers with their questions to predict whether the answer will get accepted or not. We perform an extensive empirical analysis on the retrieved dataset to identify and extract features and perform correlations between various variables to know the causality. Our findings will suggest that:

- Prior involvement of the answerer on question tags and topics increases the chance to give the answer for that question.
- Expertise will increase the chance in acceptance of the answer.
- Topical compatibility between the question and answer increases the satisfaction of asker or community with that answer.

Outcome

Armed with this observation, we have used classification algorithms to predict acceptability of the answer by the asker or community.So,the outcome of this study lies around predicting the acceptance of the answer as the best answer and various other performance metrics and accuracy ratios that deal with it.

References

Yuhua Lin, Haiying Shen (2015)
SmartQ: A Question and Answer System for Supplying High-Quality and Trustworthy Answers
Journal of Latex Class Files Vol. 14, NO. 8, August 2015.

Tirath Prasad Sahu, Naresh Kumar Nagwani, Shrish Verma (2016)
Selecting Best Answer: An Empirical Analysis on Community Question Answering Sites
Date of publication : August 26 , 2016
Digital Object Identifier 10.1109/ACCESS.2016.2600622

The End