CS 6350 Machine Learning Project Final Report

Predict Closed Questions on Stack Overflow.

1 Team Members

- Aravind Senguttuvan (u0877727)
- Anirudh Narasimhamurthy (u0941400)

2 Description of the problem

The problem which we worked on involves building a classifier which predicts whether or not a question posted on StackOverflow will be closed, given the question as submitted. Broadly speaking our problem is a multi-class classification problem and the labels which the classifier outputs are:

1. Open
2. Off Topic(OT)
3. Not Constructive (NC)
4. Not a Real Question(NRQ)
5. Too Localized(TL)

If the labels corresponds to 1 it is an indication that question is most likely to be open. If the labels correspond to any one of 2, 3, 4 or 5 then there is a high probability that the examples or questions or posts corresponding to those labels will be predicted to be closed.

3 Interestingness of the problem

Given that StackOverflow is a question answer service which is widely used by programmers and technical personell all across the world, it requires more resources to moderate and maintain its standards. An automation of the process of closing questions on StackOverflow would be a small step in reducing the

The problem statement was provided in a contest organized by Kaggle, but as people who have been exposed to Machine Learning, this problem has several interesting components which could be solved using the concepts and algorithms we have learnt in the course. Another interesting feature of the problem is to find a way to convert the multi-class classification problem finally to a binary classfication problem which would instruct or intimate someone if a question needs to be closed or not. If the model works accurately, then there wouldn't be a need of manual intervention to close a question posted on StackOverflow.

4 Dataset explained

Kaggle had provided two different datasets: private and public dataset against which the users could test their predictions. We ran our experiments on the public training data.

Number of examples:

Label	Number of examples
Open	89337
Not a Real Question	38622
Not Constructive	20897
Off Topic	20865
Too Localized	8910

Table 1: Distribution of training examples for each label

Not all questions submitted to StackOverflow are valid and few questions are flagged invalid and closed due to one of the following reasons: Off topic, not constructive, not a real question, too localized or exact duplicate. Closing a question involves voting by other users and once it reaches a specific count, the moderator then closes the question. This involves a lot of work.

To make life easier and to have a sophisticated software solution for moderation StackOverflow posted its dataset on a Kaggle competition to identify the best prediction model for closing questions.

5 Important Ideas Explored

6 Ideas from class which were used

7 Description of the techniques/algorithms used

8 Lessons Learnt/Observations

9 Results

10 Future work

11 Techniques Used

The paper `http://ceur-ws.org/Vol-1031/paper2.pdf` uses Random Forest, Support Vector Machine(SVM) and Vowpal Wabbit techniques to determine if a question should be closed. We are trying to prove the paper's evaluation.