WHY WILL MY QUESTION BE CLOSED? NLP-BASED PRE-SUBMISSION PREDICTIONS OF QUESTION CLOSING REASONS ON STACK OVERFLOW

Tushar Pandurang Kadam

Indian Institute of Science

ABSTRACT

Stack Overflow is a popular Community-based Question Answer website focused on software programming and has attracted more and more users in recent years. As of October 2021, Stack Overflow had more than 22 million questions. Although the posting ethics were guided in detail, the quality of many posted questions is poor [1]. A question on SO is closed if it doesn't follow guidelines from SO. But often reason for closing particular question is blurred to user, which leads to debates and occasional negative behavior in answers or comments. With the aim of helping the users compose good quality questions, Toth et al proposed [2] a set of classifiers for the categorization of Stack Overflow posts prior to their actual submission. In this work, I have implemented their [2] **architecture from scratch** and tried to **replicate their results**. I have also provided additional visualizations of **intermediate representations** of architecture to verify if architecture is learning correct representation of Stack Overflow posts.

1. INTRODUCTION

Stack overflow is a huge community driven question answering website. As of October 2021, It has 22 million questions, 32 million answers, 83 million comments and 62 thousand tags. Whether you are professional or enthusiast programmer certainly Stack Overflow is a very popular platform. SO's immense popularity clearly shows the ever-growing impact of social media in Software Engineering, which, however, has its certain drawbacks. In particular, the competition between quantity and quality of questions is increasing as the number of posts rises, which leads to issues in maintaining the professionalism of the site. Given the extreme posting frequency, moderator work is undoubtedly an elaborate and laborious task. At the same time, moderation is the key to avoid the issue of quality decline, which is highly important for the sustainability of the service SO provides. In order to maintain a reasonable level of quality, supporting poor questions should be avoided, and closing or deleting them is often inevitable. Nevertheless, the definite reason for closing a particular post is often unclear to the users, which in frequent cases leads to heated debates. Essentially anyone with any experience level can ask a question on SO but there are a few community-established rules for posting. Questions not satisfying these requirements will be closed or deleted by moderators or privileged users. Currently, there are five reasons for closing a question:

- Due to duplication.
- The question is off-topic.
- Unclear what the user is asking.
- The question is too broad and cannot be answered straightforwardly.
- The question is primarily opinion-based leading to subjective discussions.

The closing procedure is a manual task relying on a voting system. Moderators or privileged users can cast a vote if they find a question inappropriate, and five such votes would end up in closing the question. Previous studies in this topic essentially focused on the classification of SO questions according to their quality relying on features extracted primarily from **user data or post-submission** information available only after community feedback. The most important feature among those is reputation / experiance of user because more reputed question posters are least likely to post question that may be closed because of their experiance.

On contrast authors of this architecture wants to classify questions into the existing closed categories of SO prior to submission, that is, in a form visible only for the posting user. This pre-evaluation tool would assist the users in composing a question, which will very likely remain open and receive responses.

In order to solve this problem authors tried to implement **Gated-Recurrent-Unit-based (GRU)** classifier augmented with Natural Language Processing (NLP) tools capable of predicting whether a question will be closed by relying only on the textual features of the post. This is the first work to classify SO questions and also classify exact reason for closing of the questions.

Many times its not clear by which reason question was closed hence we can see following comments discussions regularly. [1]

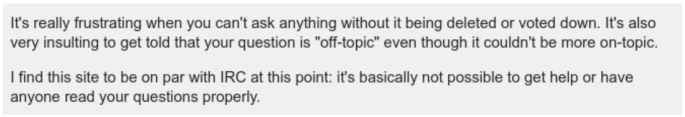

Fig. 1. Frustration for closing and deleting a question

The objective of authors work is to provide the user with a practical tool capable of pre-evaluating the questions to be posted on SO, and determining whether the question will be marked for closure by the community after submission. With these goals in mind, contribution contains three essential ideas:

- Classifiers relying exclusively on the textual properties of a question: its title, body
- Not using any post-submission information, such as number of answers, scores, or user reputation.
- Directly predict the exact closing reason of a question.

[1] https:/meta.stackoverflow.com/questions/388076/is-this-how-we-want-to-treat-newcomers

2. METHODS

Figure 2 shows overview of architecture that author has used. Details will be explained later.]

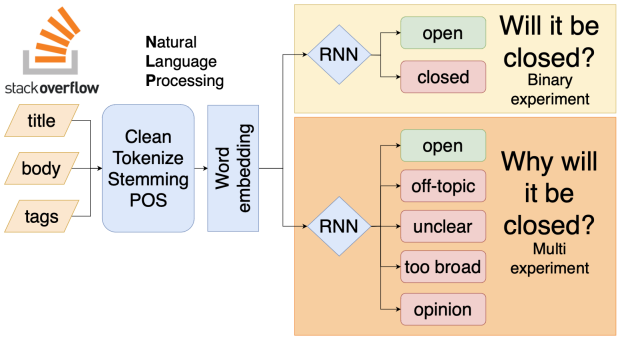

Fig. 2. Model Architecture

Code for this work can be viewed in Github repository [2]

2.1. Dataset Preparation:

The dataset used is Stackoverflow data dump till Sept 2021 [3]. Removed questions posted before June 2013 because after that current question closing policy was introduced. Questions closed due to duplication were also excluded from our dataset because of two reasons: i) this problem has already been well examined previously [3]; and ii) finding a duplicate would imply the knowledge of the previously posted questions, which is the opposite of the purpose of using only pre-submission information.

Stackoverflow has humongous dataset - 5.5 crore posts. There are around 1 million closed posts. In compressed format its around 30GB and expands to 401 GB SQL server database.

Following steps were used to prepare Dataset for computationally feasible data set:

- `Posts.xml` from SO data dump contains all the posts from Stack Overflow in XML format. (around 18 GB)
- Using XML parsing libraries converted this XML Dataset to required CSV format with the help of streaming API's because its difficult to load complete `Posts.xml` once in memory - `stackexchangeparser` folder in Repository
- Questions which are closed have `ClosedDate` attribute as non null.
- Sampled all Closed questions and sampled 2% of Open questions: which have nearly same number as total Closed questions - `sample.ipyth` Notebook
- Performed EDA and segregated `Open` and `Closed` questions from CSV - `EDA.py`
- To obtain reason for Closing Question there is another XML file in Stackoverflow Data dump : `PostHistory.xml` which records every event on posts.
- Stack Exchange networks have Data Explorer[4] which is web interface of SQL Server of every Stack Exchange data dump. Queries can be fired and statistics can be calculated on real SE dataset. If results are small we can download them in CSV file hence I downloaded all the closed questions and their reason of closing questions using this method instead of downloading complete `PostHistory.xml` (27GBs!).
- Performed joins on tables from previous step and set of all closed questions obtained from `Posts.xml` to get reasons for closing that questions - `EDA.py`
- Removed Duplicate questions - `EDA.py`

These were the steps performed to provide to obtain SO dataset. In final dataset there are 584051 posts among which 265163 are closed. Class distributions are as follows, Open: 54.5%, Off-Topic: 20.7%, unclear: 10.7% , Too-broad: 10.11% and Opinion: 3%.

2.2. Input preparation for Model:

After removing Nulls and replacing NaNs with empty strings. I concatenated `Title`, `Body` of SO post to form a single a column of `title_body`. Using `BertTokenizer`.[5] converted text data into tokens and which in turn are converted to tensors in `csv2tensor.py`

2.3. Model:

In `models` directory there are `GRU.py` and `Bi_GRU.py` two files which contains code for models. First one uses unidirectional GRU layer while second one uses bidirectional GRU layer. In both the models first layer is Embedding layer then GRU layer and finally output from each time step of GRU layer is concatenated and passed it to fully connected layer for classification task. I have added 2 layers (stack) of RNN's in this task. Simple GRU layered model has around **10 million** parameters.

2.4. Hyper Parameters of model:

`Modules.py` contains list of all the parameters and their values at one place.

2.5. Evaluation and Metrics:

`evaluate.py` contains code for model evaluations where predictions are generated and saved. Later `metrics.py` is used to generated various metrics from these predictions like precision, recall and accuracy.

3. EXPERIMENTS:

Binary classification and Multi class classification experiments were conducted on SO datasets in this work.The second one is a five-class classification predicting the different closing reasons (off-topic, unclear, too broad, opinion-based) versus the open state. topic, unclear, too broad, opinion-based) versus the open

[2]https://github.com/Kadam-Tushar/Why-Will-My-Question-Be-Closed
[3]https://archive.org/details/stackexchange
[4]https://data.stackexchange.com/
[5]https://huggingface.co/transformers/fast_tokenizers.html

state. Two different recurrent neural network (RNN) models denoted as UNI, BI GRU's representing Unidirectional and Bidirectional GRU were employed for this task. All networks apply the Adam optimizer together with cross-entropy as loss function.For the evaluation metrics, we applied micro and macro averages of the precision and recall pairs;

The input data was split into train and test sets applying the stratified k-fold cross-validation strategy with k = 3 yielding 3 different train and test sets.

4. RESULTS:

Average performance for the binary classification experiment

Metric	UNI	BID
Micro precision	74.35%	74.90%
Micro recall	74.39 %	74.90 %
Micro F1	74.37 %	74.90 %
Macro precision	74.86 %	76.85 %
Macro recall	73.54 %	74.62 %
Macro F1	74.20 %	75.73 %
Accuracy	**74.35%**	**74.90%**

The evaluation metrics of the binary classification experiment are shown in above figure. The table presents the averages over the results obtained for the test sets in the 3-fold calculations.

The number of related studies in the literature is scarce which uses only textual information of stack overflow questions. There is a recently published study that also used textual information exclusively, however, the classification task was different from that of this work. Tóth and co-workers[4] focused on question quality and obtained an accuracy of 74% for classifying SO posts in good versus bad categories using a deep neural network approach.

To visualise the effectiveness of learned SO post embedding, PCA and t-SNE plots are generated with different coloring for each class of post.

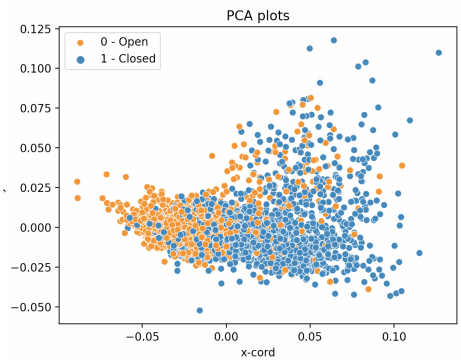

Fig. 3. PCA plot

From the plots we can observe for the binary classification problem these two classes are separable in 2 dimensions.

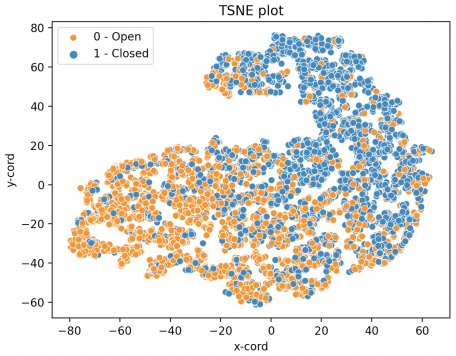

Fig. 4. t-SNE plot

Average performance for the five-class experiment

Metric	UNI
Micro precision	63.48%
Micro recall	63.43%
Micro F1	63.48%
Macro precision	51.42%
Macro recall	48.27%
Macro F1	48.47%
Accuracy	63.48%

(1)

In this second experiment we try to answer why the stack overflow question was closed. So basically this five-class classifier using the same pre-submission textual information as the binary model above is designed with the labels off-topic, unclear, too-broad, opinion-based, and open. The original dataset is imbalanced in terms of these 5 classes.

The confusion matrix of test set is as follows:

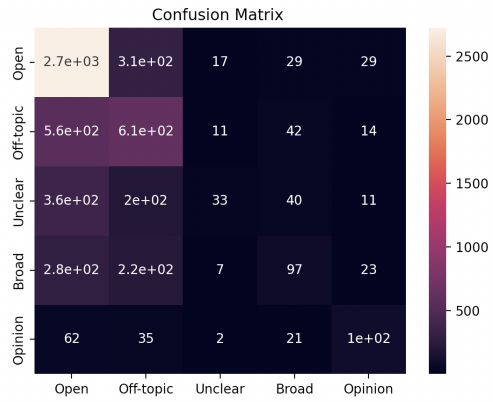

Fig. 5. Confusion matrix of Multi-class experiment

As we can see from confusion matrix this 5-class classifier is much better than random guessing. Even though there is huge class imbalance between classes model is able predict most of the examples from class 0 - open. Most of the open questions are correctly classified but model is performing worst on classes 2,3 i.e Unclear and Broad questions.

To demonstrate class imbalance we can see confusion matrix with their relative proportion. Among around 50% samples of open questions model can predict 46.6% correctly but on the other classes it is performing badly. These observations hints us to apply class balancing techniques and re-train our model.

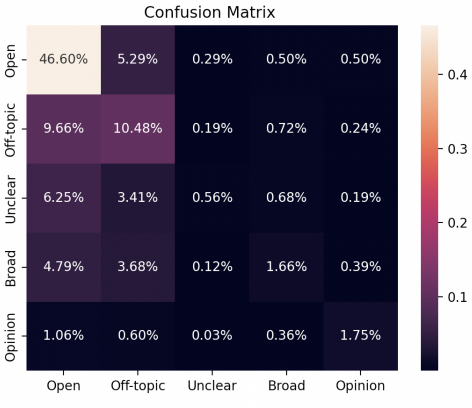

Fig. 6. Confusion matrix - Class proportion

To visualise the effectiveness of learned SO post embedding PCA and t-SNE plots for 5-class classification are as follows:

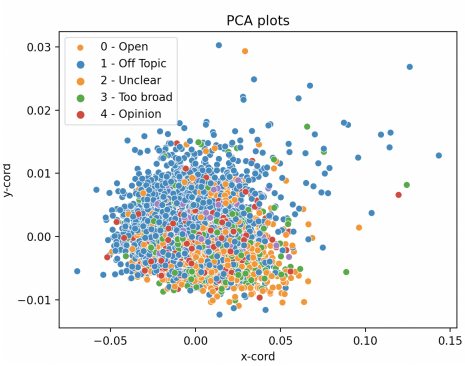

Fig. 7. PCA plot of 5-classes

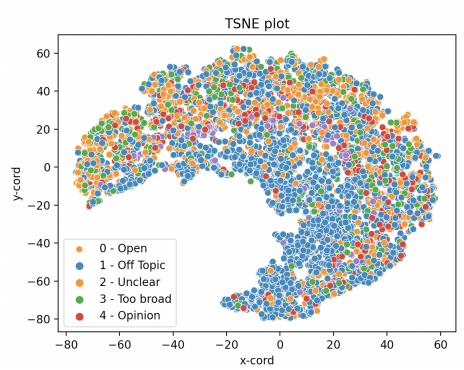

Fig. 8. t-SNE plot of 5-classes

5. CONTRIBUTIONS

- **Pre-processing**: Tokenization, Sampling, Dataset generation.
- **Developing architecture from Scratch** Developed this GRU architecture to work with word embedding from scratch
- **Visualizations**: 2D plots of vector representation of SO questions using t-SNE and PCA.
- **Model Evaluation**: Tried precision, recall, F1-score.

6. DEVIATIONS FROM ORIGINAL WORK

:

- Original work used exactly balanced classes for binary classifier but in this work I used 55-45 distribution.
- Tags and POS - part of speech tagging is not used in inputs.
- Original work used dataset till 2019 but this work has considered dataset till Sept-2021.
- Used hidden states from all the timesteps to better capture long term dependency but on the other its not clear from paper what exactly they have used in their work for e.g only last timestep hidden state or every hidden state.
- Because of time required for training I was not able to train Composite model (3rd model) from paper. But author mentioned best results were only obtained from Unidirectional simple GRU model hence given priority to implement that.
- Author used 30-fold cross validation for more reliable statistical analysis but I used 3-fold and 6-fold because of interest of time.
- As for k-fold cross validation I used small k hence I was not able to run statistical tests like KS and ANOVA for proving statistical significance for results.

7. REFERENCES

[1] Denzil Correa and Ashish Sureka, "Chaff from the wheat: Characterization and modeling of deleted questions on stack overflow," in *Proceedings of the 23rd International Conference on World Wide Web*, New York, NY, USA, 2014, WWW '14, p. 631–642, Association for Computing Machinery.

[2] László Tóth, Balázs Nagy, Tibor Gyimóthy, and László Vidács, "Why will my question be closed? nlp-based pre-submission predictions of question closing reasons on stack overflow," in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results*, New York, NY, USA, 2020, ICSE-NIER '20, p. 45–48, Association for Computing Machinery.

[3] Rodrigo F. G. Silva, Klérisson Paixão, and Marcelo de Almeida Maia, "Duplicate question detection in stack overflow: A reproducibility study," in *2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 2018, pp. 572–581.

[4] László Tóth, Balázs Nagy, Dávid Janthó, László Vidács, and Tibor Gyimóthy, "Towards an accurate prediction of the question quality on stack overflow using a deep-learning-based nlp approach," in *Proceedings of the 14th International Conference on Software Technologies*, Setubal, PRT, 2019, ICSOFT 2019, p. 631–639, SCITEPRESS - Science and Technology Publications, Lda.