Classifying Question Quality of StackOverflow

MIDS 266 - Fall 2022

Paco Valdez

The Problem

There are multiple forums where coding questions can be asked, an interesting problem is to be able to prioritize “good” questions and do the opposite to “bad” questions. The applications to this solution would help these forums to provide feedback to users while they are asking the question, also “good” questions could be moved to the next step of the workflow that could be an automated tagging system or to a manual review.

Previous Efforts

- Piyush Arora, Debasis Ganguly, and Gareth J.F. Jones. 2016. **Nearest Neighbour based Transformation Functions for Text Classification: A Case Study with StackOverflow.** In Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (ICTIR '16). Association for Computing Machinery, New York, NY, USA, 299–302. https://doi.org/10.1145/2970398.2970426
- László Tóth, Balázs Nagy, Tibor Gyimóthy, and László Vidács. 2020. **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 (ICSE-NIER '20). Association for Computing Machinery, New York, NY, USA, 45–48. https://doi.org/10.1145/3377816.3381733
- Antoaneta Baltadzhieva, Grzegorz Chrupała. 2015. **Predicting the Quality of Questions on Stackoverflow.** Proceedings of Recent Advances in Natural Language Processing, pages 32–40, Hissar, Bulgaria, Sep 7–9 2015. https://aclanthology.org/R15-1005.pdf

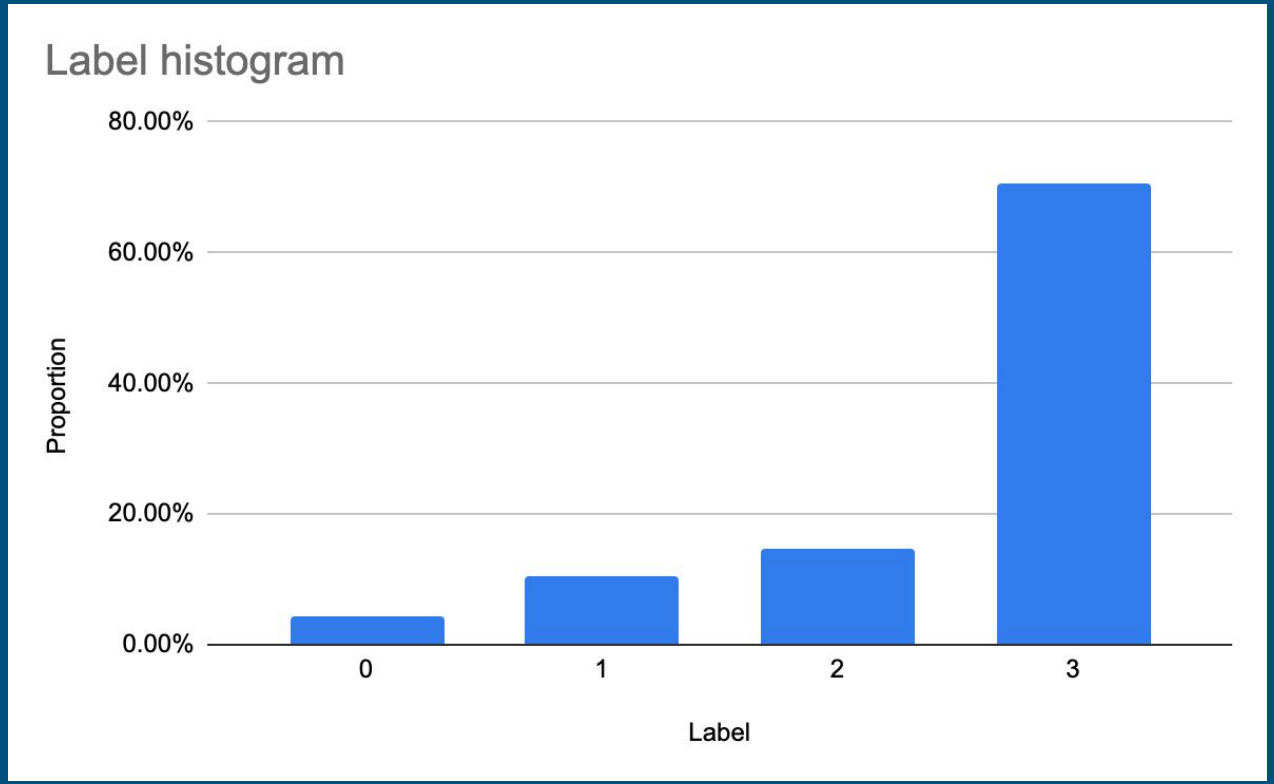
The dataset

Stack Overflow releases all the questions that have been asked and also it annotates the questions with badges. This can be translated into the following labels to be used for training.

Score-Based	View-Based	Bookmark-Based	Label
Great Question (Score >= 100)	Famous Question (Views>=10,000)	Stellar Question (Bookmarks>=100)	0
Good Question (Score >= 25)	Notable Question (Views>=2,500)	Favorite Question (Bookmarks>=25)	1
Nice Question (Score >= 10)	Popular Question (Views>=1,000)	*Missing Category (Bookmarks>=10)	2
No Badge	No Badge	No Badge	3

EDA - Label distribution

Label Distribution
Great - 0 - 4.25%
Good - 1 - 10.43%
Nice - 2 - 14.85%
Bad - 3 - 70.47%



The dataset - How questions are represented?

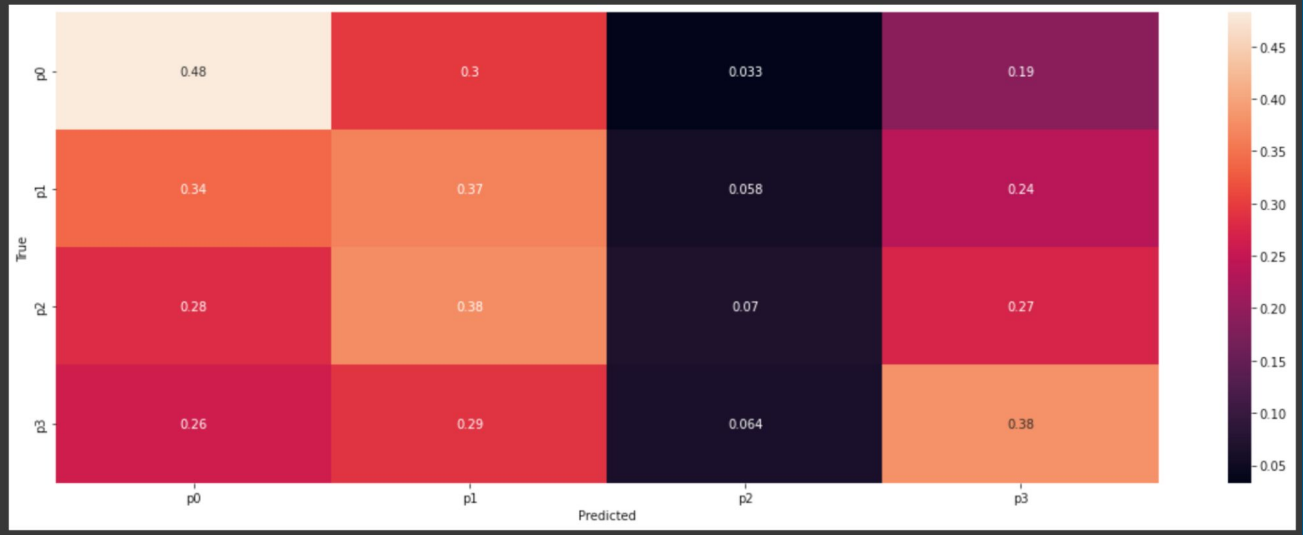
Title: Error with Redshift jdbc connector
Body: <p>Hey, I have this code to connect to Redshift using jdbc<p><code>
var jdbc = new (require('jdbc'));
var config = {
  libpath: 'C:/Users/ABCD/Desktop/jar/RedshiftJDBC41-1.1.6.1006.jar',
  drivername: 'com.amazon.redshift.jdbc41.Driver',
  url: 'jdbc:redshift://examplecluster.abc123xyz789.us-west-2.redshift.amazonaws.com:5439/dev',
  user: 'xxxx',
  password: 'xxxxx'
};

jdbc.initialize(config, function(err, res) {
    if (err) {
    console.log(err);
  }
});
</code><p>The error is the following:<p><code>
^ TypeError: undefined is not a function

at JDBCConn.initialize
(C:\Users\ABCD\node_modules\jdbc\lib\jdbc.js:62:20)
</code>

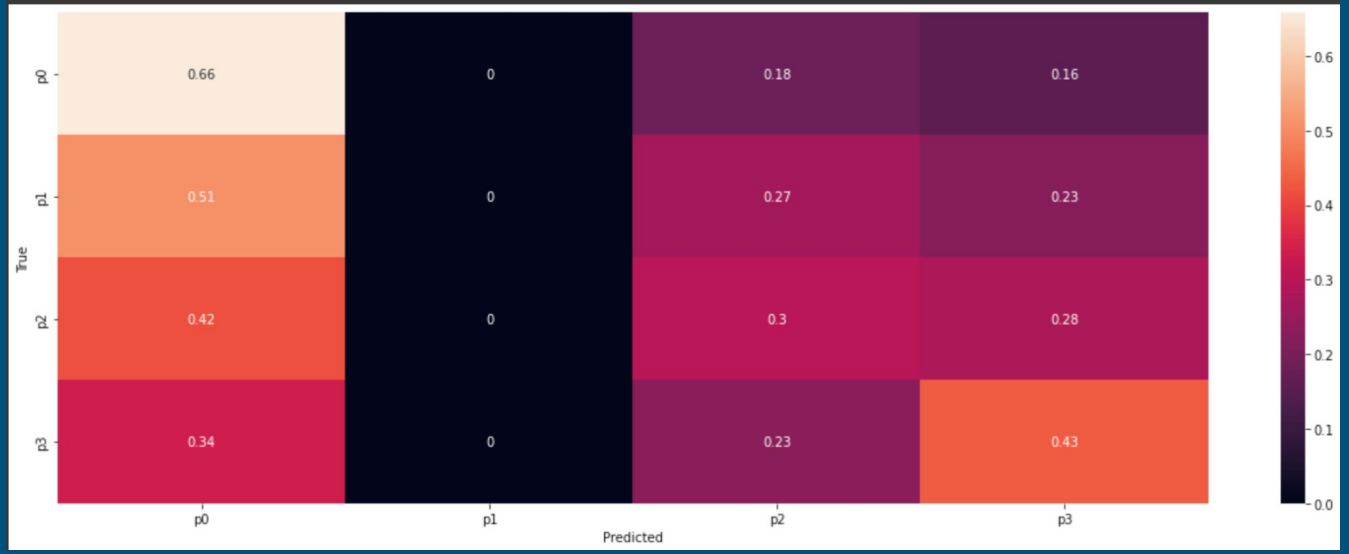
Baseline - Uncased-Bert Fine Tuned with a 20% random sample of all questions from 2012 to 2022

	precision	recall	f1-score	support
0	0.35	0.48	0.41	9849
1	0.27	0.37	0.31	9753
2	0.31	0.07	0.11	9835
3	0.35	0.38	0.37	9818
accuracy			0.32	39255
macro avg	0.32	0.32	0.30	39255
weighted avg	0.32	0.32	0.30	39255

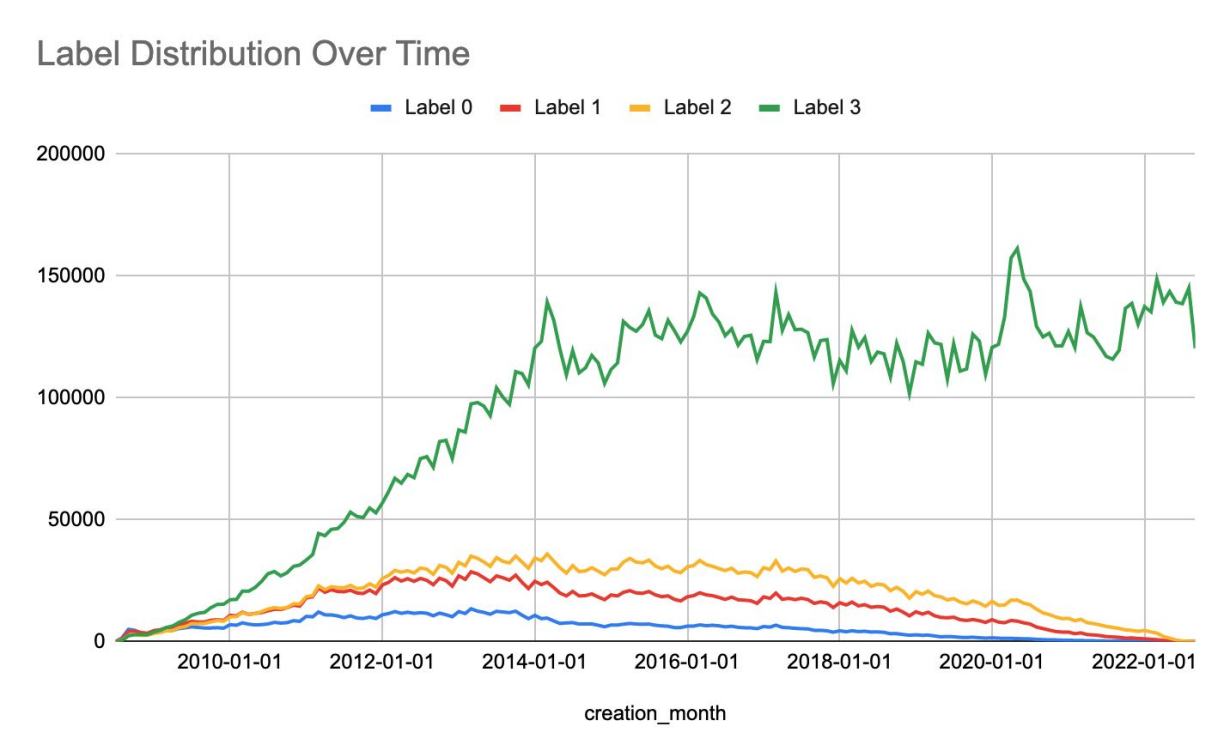


Bertweet Fine Tuned with a 20% random sample of all questions from 2012 to 2022

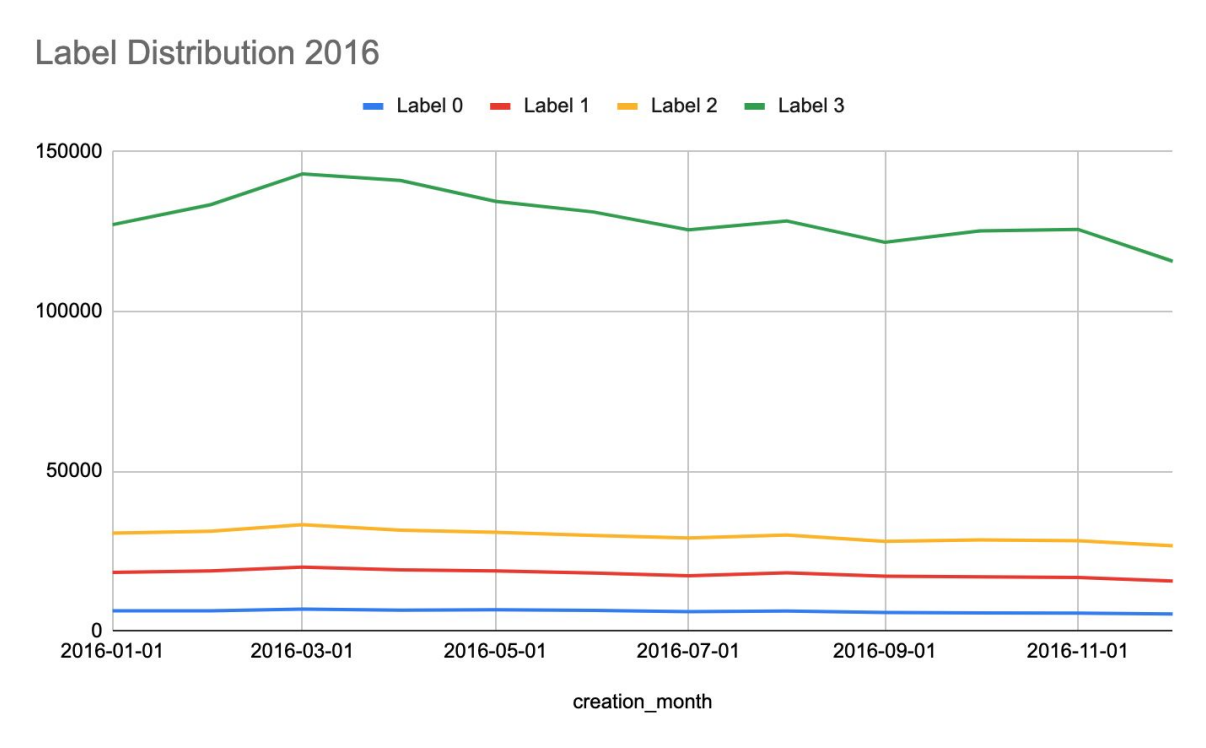
	precision	recall	f1-score	support
0	0.34	0.66	0.45	9849
1	0.00	0.00	0.00	9753
2	0.31	0.30	0.30	9835
3	0.39	0.43	0.41	9818
accuracy			0.35	39255
macro avg	0.26	0.35	0.29	39255
weighted avg	0.26	0.35	0.29	39255



Exploratory Analysis -Label distribution over time



Exploratory Analysis - Label Distribution 2016



Add special token - Unprocessed Text

Vuex getter not updating
++++++++++
I have the below getter:
<pre class="lang-js prettyprint-override">
<code>   withEarmarks: state => {
**   var count = 0;**
**   for (let l of state.laptops) {**
**   if (l.earmarks.length > 0) {**
**   count++;**
**   }**
**   }**
**   return count;**
**   }**
</code>
</pre>
And in a component, this computed property derived from that getter:
<pre class="lang-js prettyprint-override">
<code>   withEarmarks() { return this.$store.getters.withEarmarks; },
</code>
</pre>
The value returned is correct, until I change an element within the laptops array, and then the getter doesn't update.

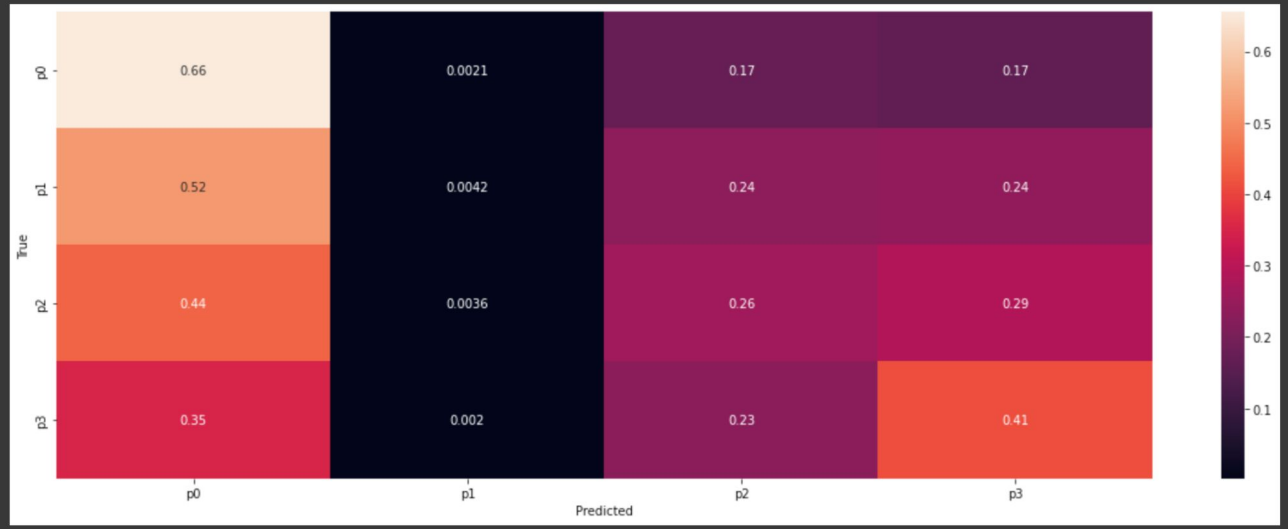
Add special token - processed Text

vuex getter not updating i have the below getter [CODE] and in a component this computed property derived from that getter [CODE] the value returned is correct until i change an element within the laptops array and then the getter doesnt update

```
from transformers import AutoTokenizer, TFAutoModel


tokenizer = AutoTokenizer.from_pretrained(model_checkpoint)
model = TFAutoModel.from_pretrained(model_checkpoint)

tokenizer.add_tokens(['[CODE]'], special_tokens=True)

model.resize_token_embeddings(len(tokenizer))
```

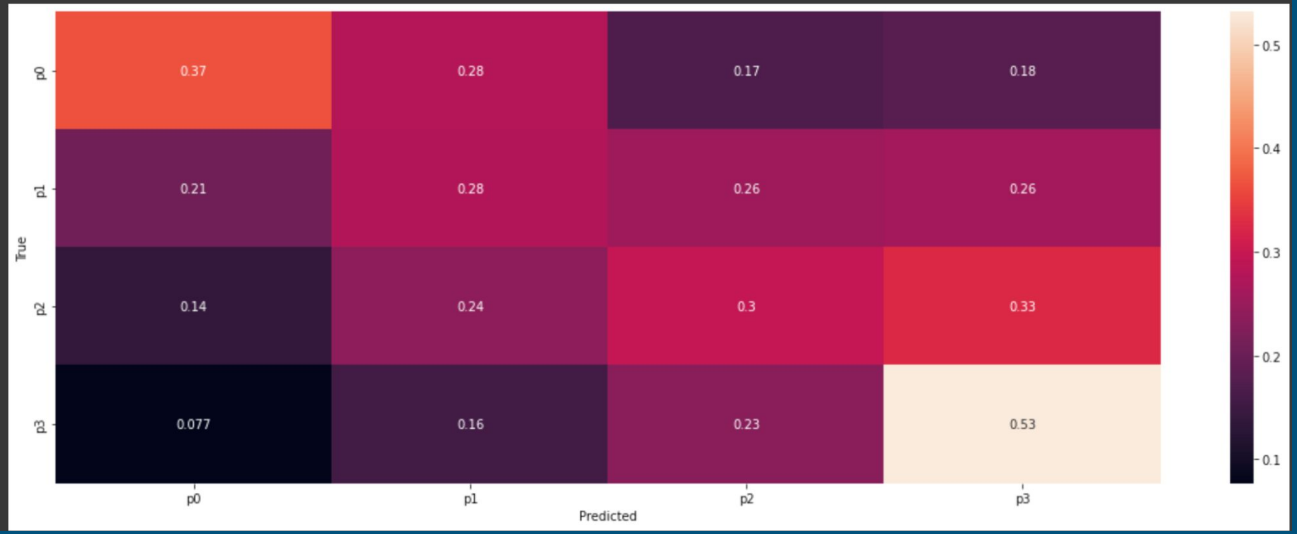
Uncased Bert with Special Token Fine Tuned with 2016 data

	precision	recall	f1-score	support
0	0.34	0.66	0.44	14988
1	0.35	0.00	0.01	14850
2	0.29	0.26	0.28	14808
3	0.37	0.41	0.39	14684
accuracy			0.33	59330
macro avg	0.34	0.33	0.28	59330
weighted avg	0.34	0.33	0.28	59330



Bertweet with Special Token Fine Tuned with 2016 data

	precision	recall	f1-score	support
0	0.47	0.37	0.41	14988
1	0.29	0.28	0.28	14850
2	0.31	0.30	0.30	14808
3	0.41	0.53	0.46	14684
accuracy			0.37	59330
macro avg	0.37	0.37	0.36	59330
weighted avg	0.37	0.37	0.36	59330



Fine Tuned T5 (2016 data)

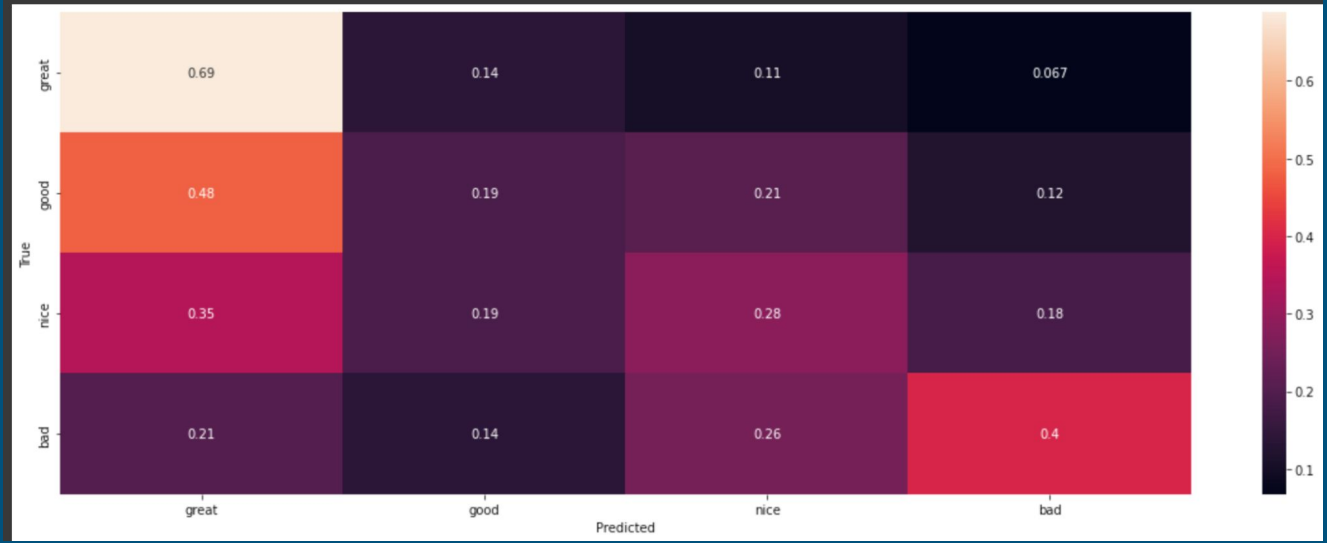
Text: I have the below getter: <pre class="lang-js prettyprint-override">
<code>   withEarmarks: state => {   var count = 0;

**   for (let l of state.laptops) {**
**   if (l.earmarks.length > 0) {**
**   count++;**
**   }**
**   }**
**   return count;**
**   }**
</code>
</pre>
And in a component, this computed property derived from that getter:
<pre class="lang-js prettyprint-override">
<code>   withEarmarks() { return this.$store.getters.withEarmarks; },
</code>
</pre>
The value returned is correct, until I change an element within the laptops array, and then the getter doesn't update.
Label: good</s>

Categories: ['great', 'good', 'nice', 'bad']

T5 Fine Tuned with 2016 data

	precision	recall	f1-score	support
0	0.40	0.69	0.51	29701
1	0.29	0.19	0.23	29729
2	0.33	0.28	0.31	29820
3	0.52	0.40	0.45	29468
accuracy			0.39	118718
macro avg	0.39	0.39	0.37	118718
weighted avg	0.38	0.39	0.37	118718



Analysis - Real Special Token Frequency

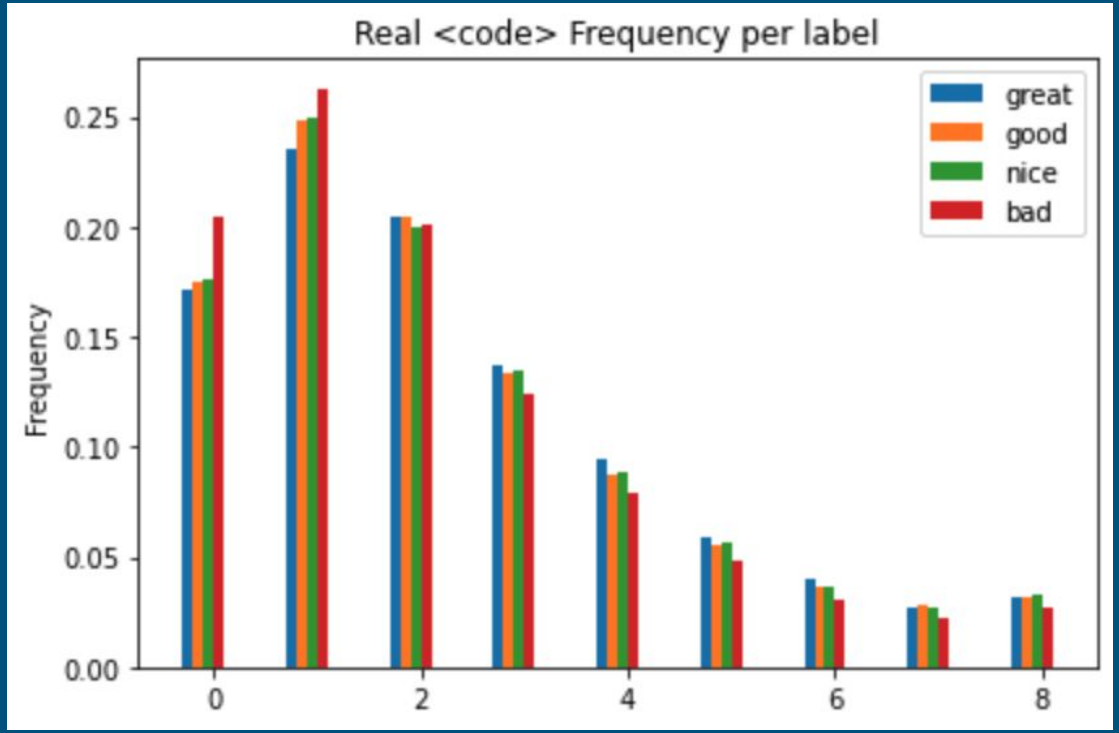
Special token frequency:

Category: **great** Mean: **2.845**

Category: **good** Mean: **2.846**

Category: **nice** Mean: **2.854**

Category: **bad** Mean: **2.557**



Analysis - Predicted Special Token Frequency

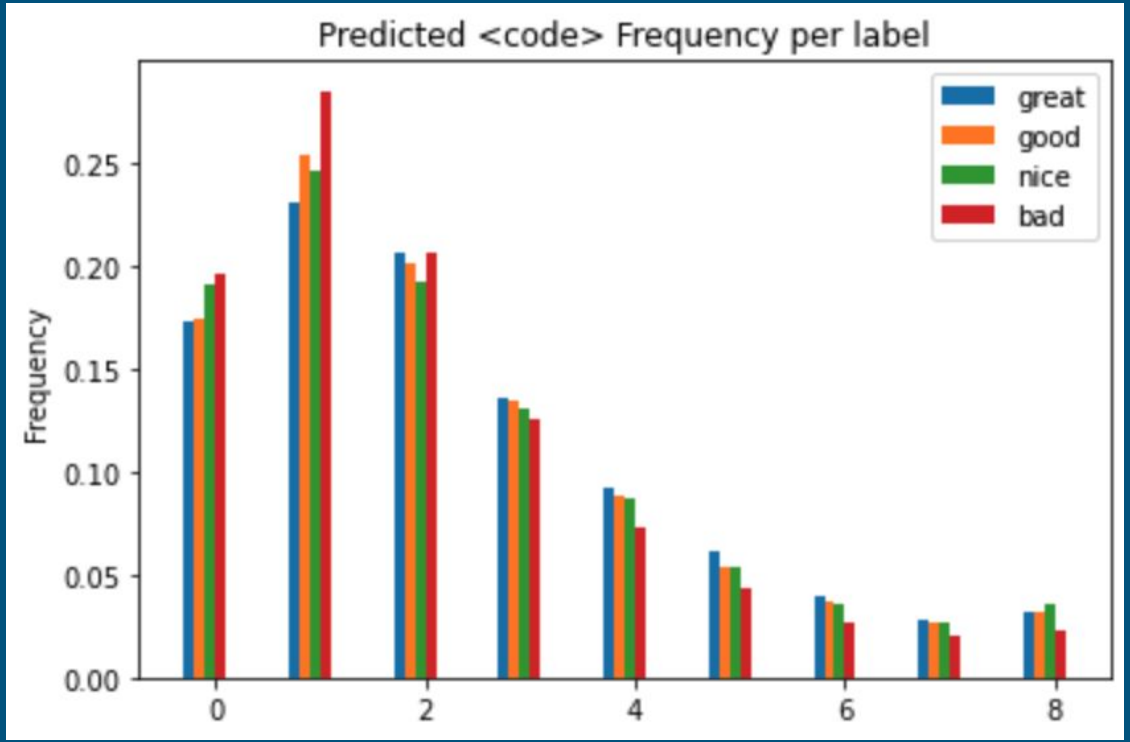
Special token frequency:

Category: **great** Mean: **2.856**

Category: **good** Mean: **2.839**

Category: **nice** Mean: **2.923**

Category: **bad** Mean: **2.376**



Next Steps

- Use only a subset of the questions. (i.e. only python questions, or exclude esoteric languages)
- Adjust the score of a question based on when was voted on.
- Use a model trained on code to analyze the code blocks.
- More complex models on top of T5 or Bertweet.
- Do binary classification.

Questions

Thank you!