**Classification of Quiz Bowl Questions**

Ryzeson Maravich

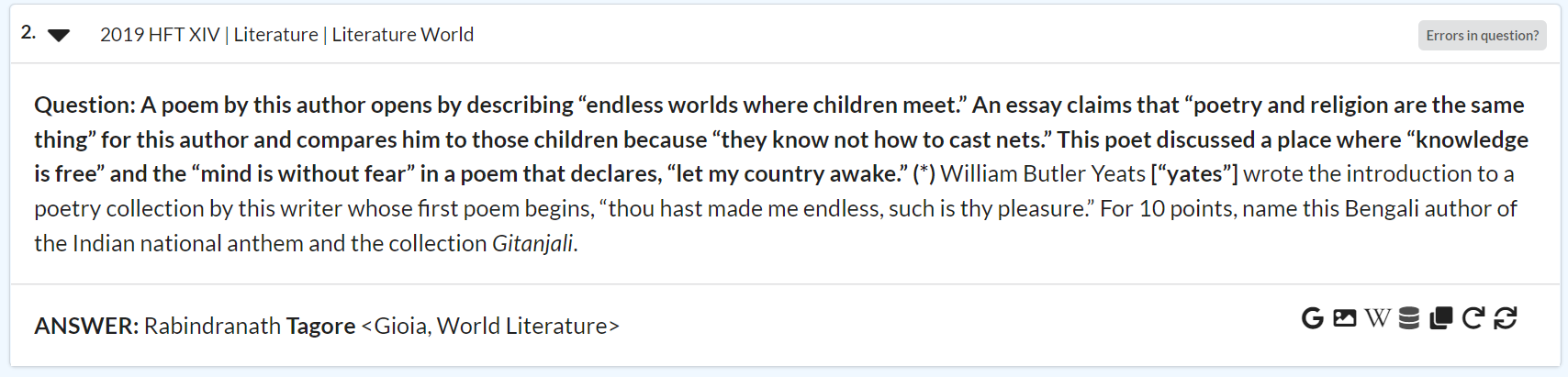
**Abstract**

This paper classifies quiz bowl questions into their appropriate category and subcategory using TfidfVectorizer and Naive Bayes classifier pipelines, as well as by computing cosine similarity. The results of these classification models are compared, and their performances are analyzed. While the performance of the subcategory classifier is suitably high, it suffers a performance drop when tested on a different trivia dataset. Finally, research limitations are addressed and potential improvements are proposed.

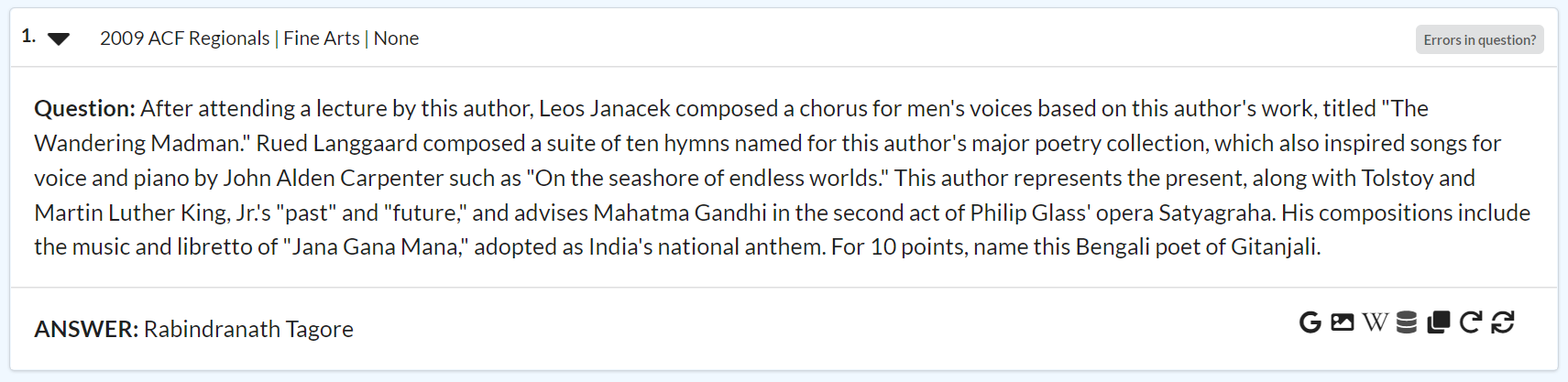
**Introduction**

This project attempts to classify quiz bowl tossups (i.e., questions that consist of about 5-8 sentences) into their respective academic disciplines using Natural Language Processing techniques. The most widely-used searchable question database in the quiz bowl community, QuizDB.org, allows users to sort by both category and subcategory. While this is a useful tool that works well most times, several questions have missing (or incorrect) category labels, and many more have missing subcategory labels. As a result, there are often times when questions that the user is trying to specifically query do not actually show up, and remain mostly hidden within this database. Additionally, any new questions added to this database must have their category and subcategory manually assigned upon entry. A successful classifier would be able to help identify currently unlabeled or incorrectly labeled questions in the database, as well as automate this assignment process for new questions.

One might think that this problem is trivial or that NLP techniques are overkill for such a simple problem. Consider the example tossup below:



A search for keywords like “author”, “poetry”, or even “Gitanjali” should be able to determine that this question should be categorized as literature. While that it is true for the question above, consider this similar question:



While these exact same words of “author”, “poetry”, and “Gitanjali” also appear in this question, it is clearly testing auditory fine arts knowledge rather than literature. This shows that question classification is far from a trivial problem, and requires a more nuanced approach that NLP provides.

**Background**

NLP is used for a variety of problems, from classifying spam to producing human-like speech. Quiz bowl questions have previously served as a corpus for projects, such as QANTA, an automated quiz bowl player (Iyyer et al.). Although that paper forgoes the “bag of words” approach that is taken here, other models have employed it to reach similar goals. Osman and Yahya use tf-idf matrices to classify exam questions based on the levels of Bloom’s taxonomy (4), which is similar in structure to classifying questions based on academic topic. Tan seems to build on this work, verifying that Naive Bayes classifiers work well for sufficiently large datasets (8), a fact that led to the choice of Naive Bayes classifiers for this study. By sharing similar model designs, there is good reason to believe that small modifications or a change/addition to our training corpus could result in the ability to classify other academic-based questions or text. In this case, the models could have additional practical benefits such as verifying that an examination has a balanced subject distribution.

**Data**

All tossup and category data was collected from QuizDB.org. Each category’s data was collected from a simple query on the website with the selected category being the only search criteria. It was then parsed from the generated json file, which includes both the tossup text, category, and subcategory information.

The following list contains all categories that the models attempt to classify, with subcategories being listed in bullets underneath their respective category. If a given category has no subcategories, then it is also treated as a subcategory. For example, while American literature is a subcategory of literature, philosophy is both a category and subcategory.

Categories:

Fine Arts

* Audio
* Other
* Visual

History

* American
* Classical
* European
* World

Literature

* American
* British
* European
* World

Science

* Biology
* Chemistry
* Math
* Physics

Geography

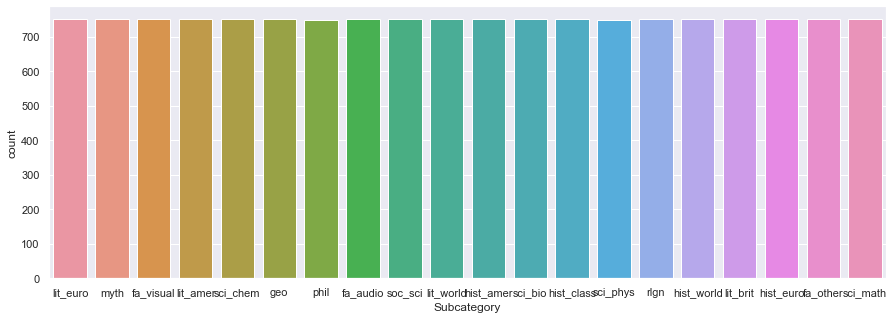
Mythology

Philosophy

Religion

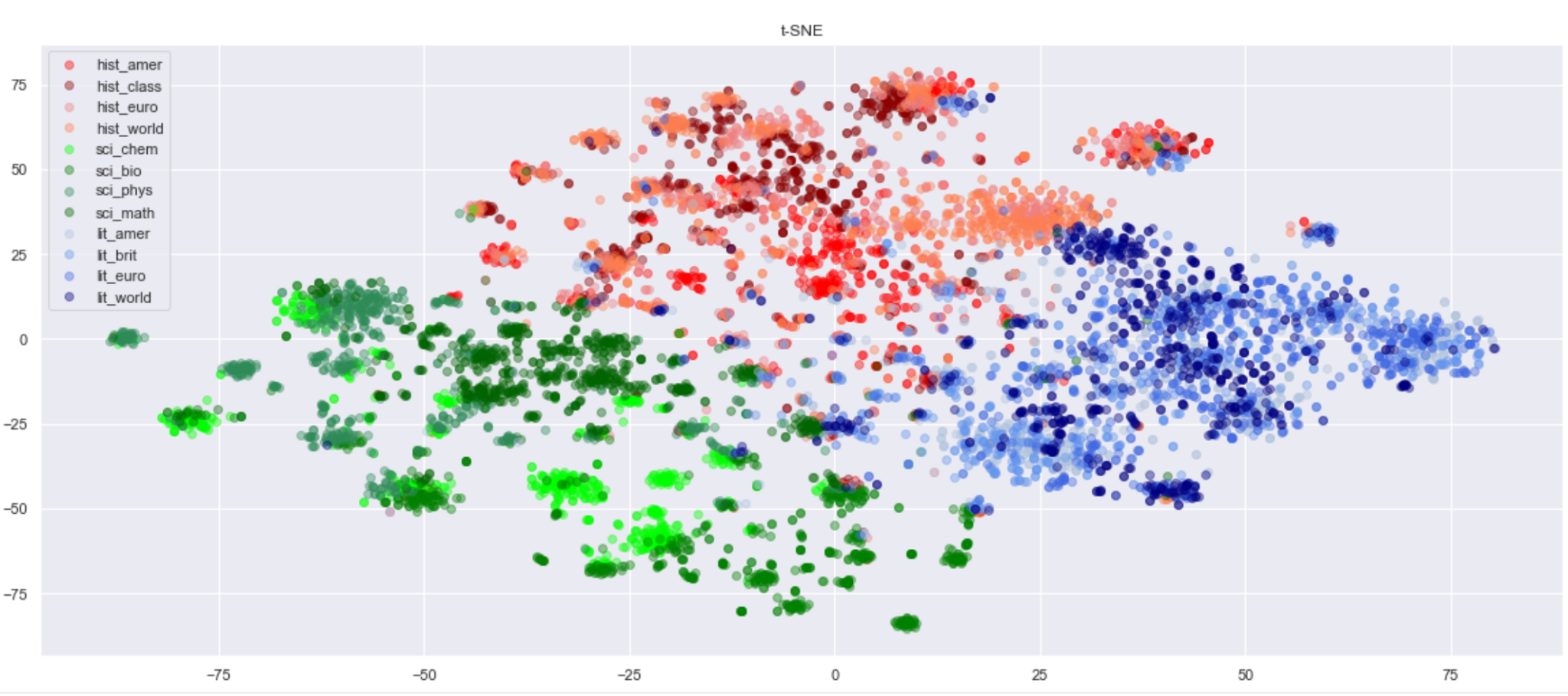
Social Science

About 750 questions were collected for each subcategory. There were occasionally questions with incomplete information that had to be discarded, but the plot below shows that our data remains balanced.



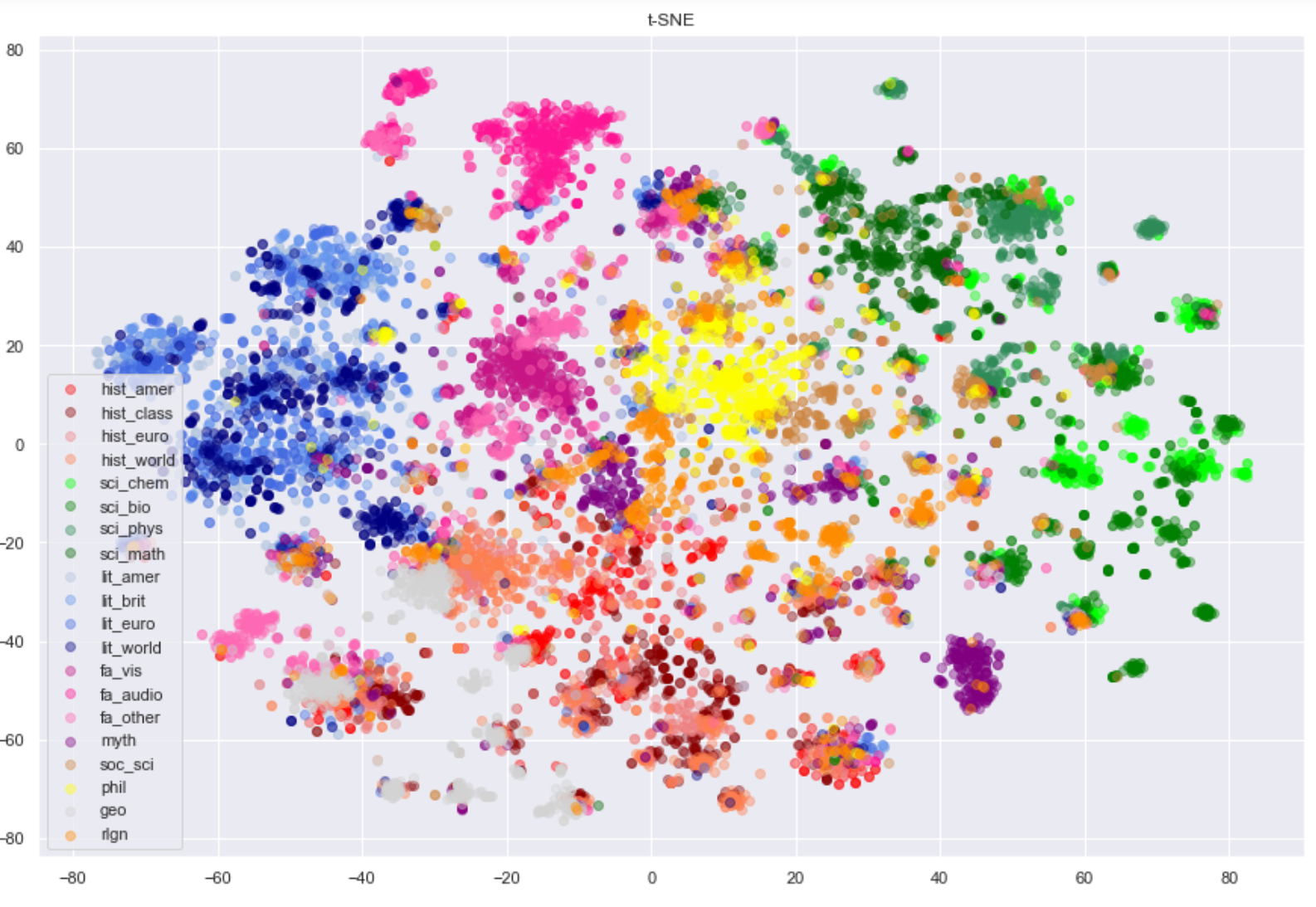
Bar chart of the number of questions in each subcategory.

After the question data was cleaned and vectorized, it was visualized using dimensional reduction. The first graph shows a 2-dimensional t-SNE projection of all vectorized history, literature, and science questions colored in shades of red, blue, and green, respectively. These “Big Three” categories were chosen because they collectively account for about 60% of all quiz bowl questions, and v. By being one of the easiest possible classification tasks, a visualization of this data should be able to provide insight into whether or not questions are able to be successfully classified at all.



2-dimensional t-SNE projection of all vectorized history, literature, and science questions.

All three categories are easily separated, occupying separate regions of this projection, and so there was good reason to believe that classification by category would work. Additionally, many of the subcategories within the broader categories are also clustered together (denoted by different shades), although to a lesser degree. Subcategory distribution is further explored with the following figure, which is a 2-dimensional t-SNE projection of questions from each subcategory.



2-dimensional t-SNE projection of vectorized questions from each subcategory.

This plot contains less compartmentalized groups of data and much more overlap between subcategories than in the previous figure, but overall the data seems mostly separable by subcategory. Especially when considering the wide range of content and diction that different subcategories likely share, the ability of this projection to successfully tease out their differences gives a promising lead that a classifier can be trained to do the same.

**Methods**

All question text was first preprocessed. This consisted of removing all stopwords and punctuation, then lowercasing, lemmatizing, and tokenizing each word. Then the data was split into 70% training data and 30% testing data, with each processed tossup counting as one data point. Three models were then created: one to classify all “Big Three” (e.g., literature, science, and history) questions by category, one to classify all questions by category, and another to classify all questions by subcategory. For each of these, a pipeline was created that vectorized the data using a TfidfVectorizer, and then trained an sklearn Multinomial Naive Bayes Classifier. Gridsesrch was used to find optimal hyperparameters for the TfidfVectorizer. Accuracy was computed simply as the percentage of testing data that the model was able to correctly classify.

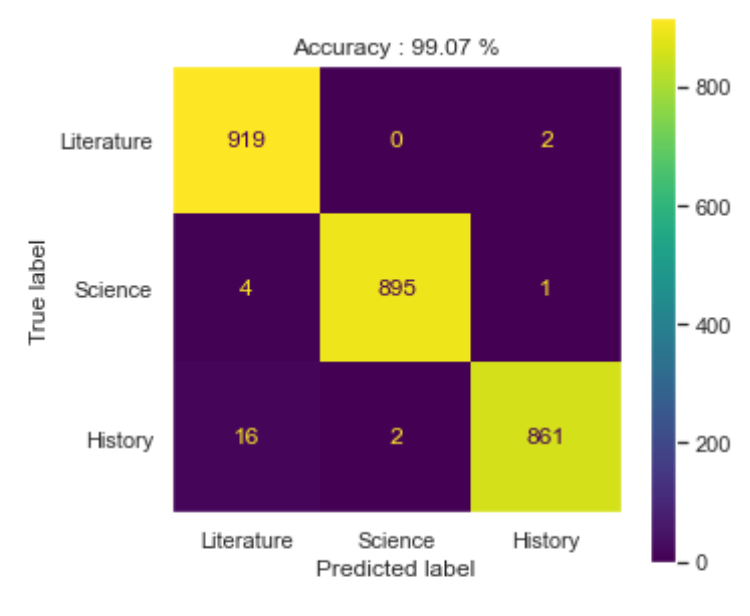
Additionally, cosine similarity was used to classify tossups based on their subcategory. For each subcategory, all training tossups were transformed using a TfidfVectorizer into a single vector. Each test data point was classified under the same subcategory as the vector for which it yielded the largest cosine similarity. Again, accuracy was computed as the percentage of testing data correctly classified.

To test how applicable the most comprehensive model was when applied to other datasets, the subcategory classifier was then tested on 750 short trivia questions found from the website TriviaBliss. This website was chosen because it had many labeled questions that matched the subcategories of the quiz bowl tossups. 50 random questions from each subcategory comprised the testing set, with the exception of the subcategories that did not have a direct analog on TriviaBliss, which included European Literature, Philosophy, Social Science, Classical History, World History, and European History.

**Results**

The table below shows the results of the three classifiers.

| **Name** | **Accuracy on test data** | **Optimal hyperparameters** |
| --- | --- | --- |
| Big Three | 99.07% | 'tfidf\_\_max\_df': 0.5 |
| All Categories | 90.04% | 'tfidf\_\_max\_df': 0.5, 'tfidf\_\_max\_features': 1000, 'tfidf\_\_min\_df': 0.001, 'tfidf\_\_ngram\_range': (1, 2) |
| Subcategories | 84.82% | default |

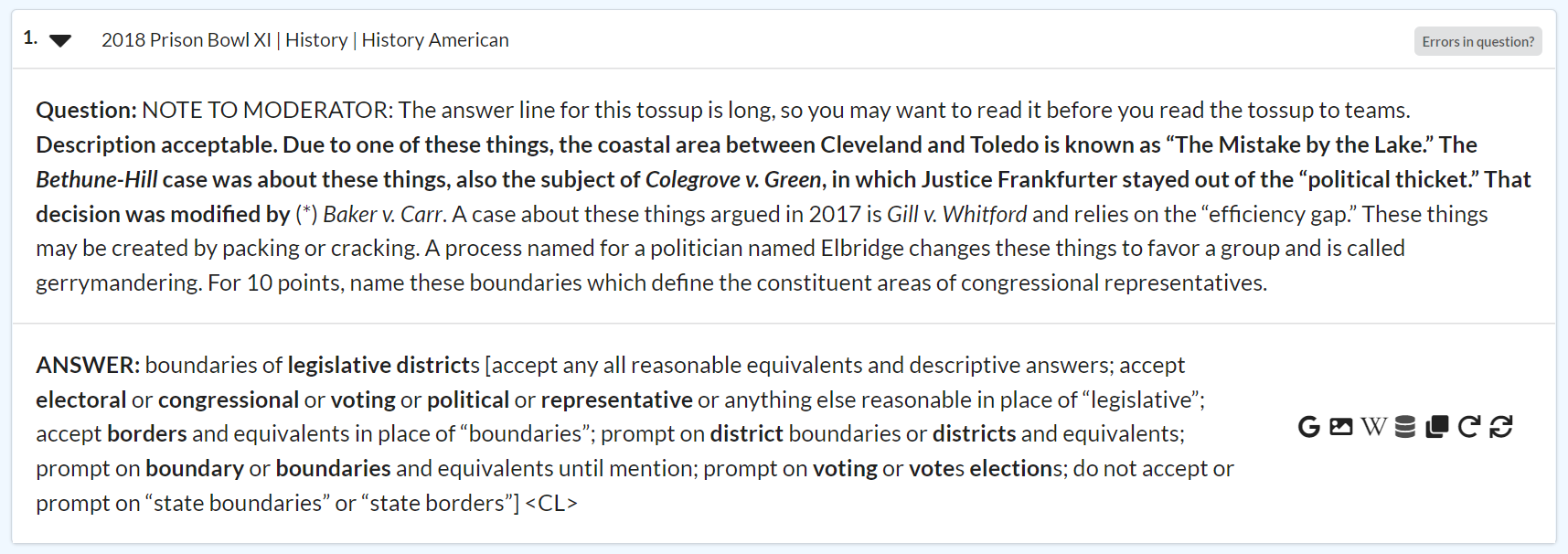
**** Confusion matrix for Big Three classifier

As expected, the Big Three classifier performed very well, achieving an accuracy of over 99%. This demonstrates that the fundamental task is able to be achieved, especially when the question categories are highly dissimilar. Taking a look at some of the few questions this classifier got wrong can help explain why it might have gotten these incorrect. Below are three questions that were classified incorrectly.

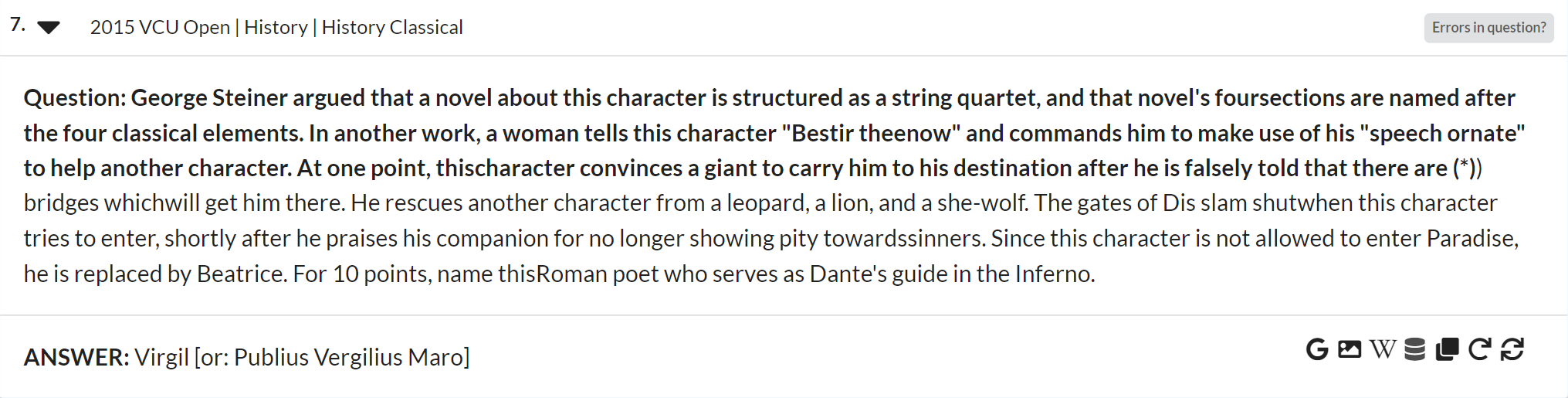
Tossup 1: A history question classified as literature.

15. 
2018 Early Fall Tournament (EFT) I History I History American 
Errors in question? 
Question: This song was written by a jeweler from Ohio named Alexander Coffman Ross, and he introduced it while on a business trip to New York. 
This song's lyrics mention the "Loco standard tottering" and claim that the "Bay State boys turned out in thousands." The lyrics of this song say to 
"let 'em talk about hard cider," a response to critics of this song's subject and attempt to portray him as a man of the people. This song was a 
prominent feature of the (*) Log Cabin Campaign. This song frequently mentions promises to "beat little Van," a reference to incumbent President 
Martin Van Buren. For 10 points, name this campaign song used by William Henry Harrison whose title references his famous nickname and his less 
famous running mate. 
ANSWER: "Tippecanoe and Tyler Too" [or "Tip and Ty"; prompt on Tippecanoe] < T R, American History> 

Tossup 2: A history question classified as science.



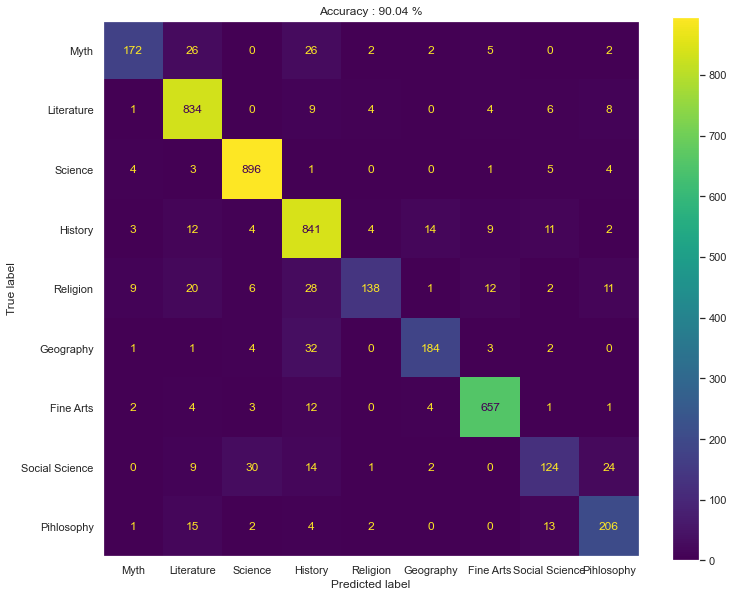
Tossup 3: A history question classified as literature.



In tossup 1, the question and answerline are obviously about history, but words like “written”, “lyrics”, and “critics” make sense in directing the classifier towards a piece of literature. Similarly, in tossup 2 there are many words that could easily appear in a science tossup such as “area”, “efficiency”, “boundaries”, and “process”. Also, tossup 2 contains a nonspecific pronoun (e.g., “this thing”) rather than an evocative one (e.g., “this poet” or “this composer”), which gives the classifier less information to work with. Another redeeming factor of the classifier is that although these are valid tossups that ask about acceptable topics, they do so in unusual ways (e.g., asking about a campaign song rather than a specific historical figure) that are not representative of most questions, so our classifier is unlikely to be misled in this way often.

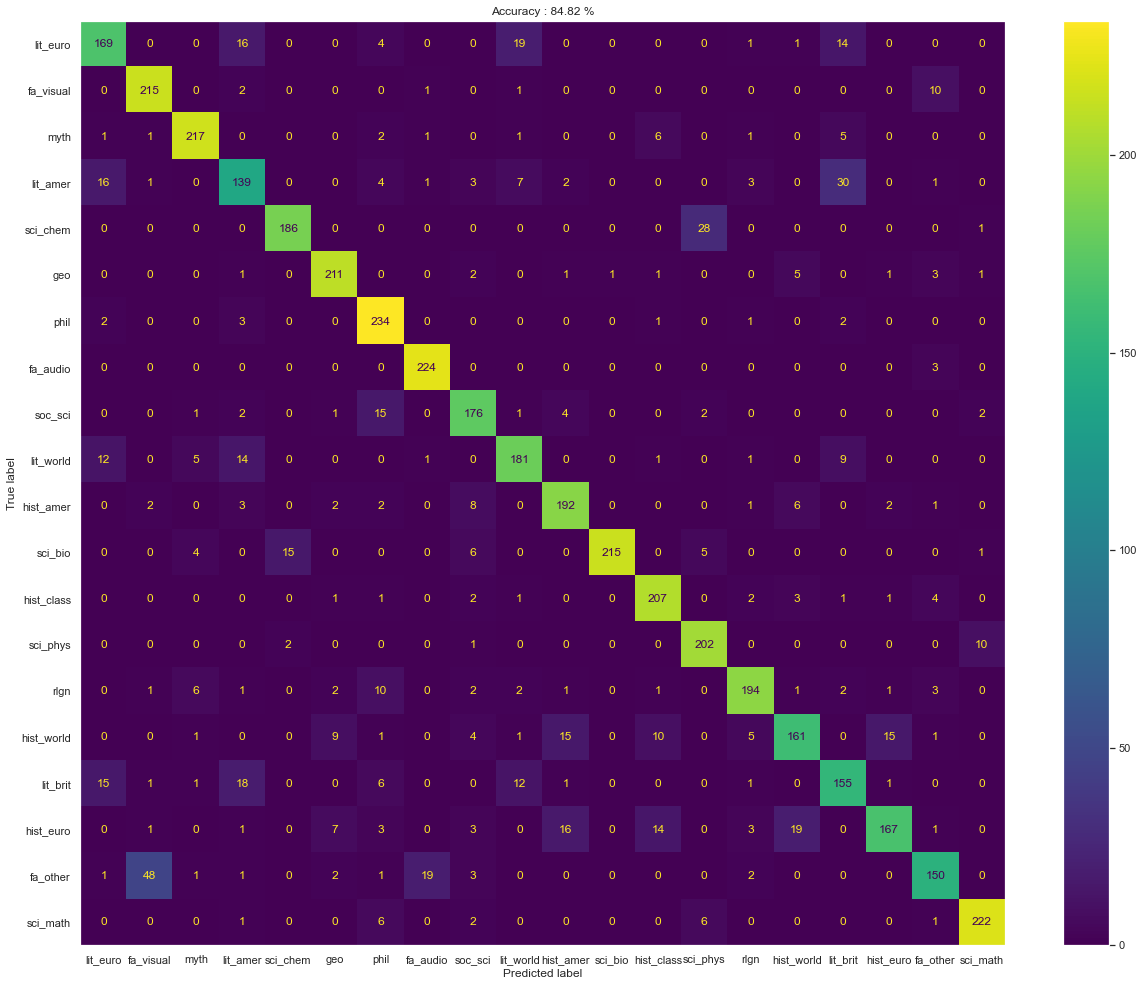
Tossup 3 was incorrectly classified as literature, although it is officially labeled as history. Upon further inspection it can be seen that the question is indeed obviously about literature, and that the history labeling in the website’s database is incorrect. This demonstrates an immediate application of this model, which was able to pinpoint an incorrectly labeled tossup among the thousands in the database, and allowing an error report to be filed to fix the mistake.

Confusion matrix for the all category classifier



When attempting to classify all categories, the model’s accuracy dropped to around 90%. This decrease makes sense as many of these new categories share vocabulary with one another, and this expansion provides the classifier with more classes for which it can incorrectly assign questions to. Still, the model achieves useful accuracy.

Confusion matrix for the subcategory classifier

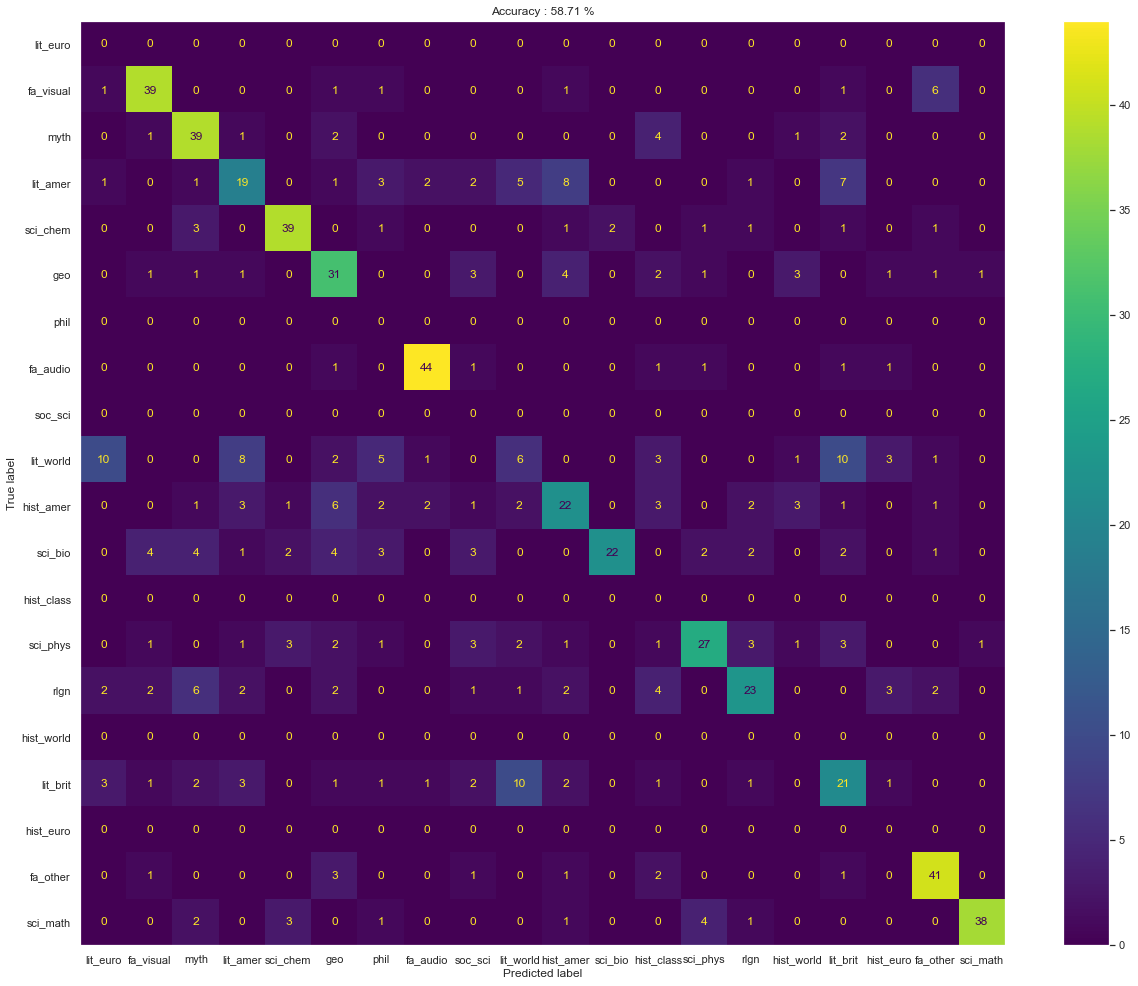


When attempting to classify by subcategory, the model again suffered an accuracy drop, this time to about 85%. Being able to further split and classify questions within a category is a much more difficult task, so maintaining this high of an accuracy is impressive. The confusion matrix shows that most of the mistakes that occur still placed a tossup into a subcategory that shared the correct overall category. For example, the confusion matrix shows that although chemistry questions were often confused for biology, they were never confused for a subcategory that was not a part of the broader category of science.

Interestingly, although the shift from the category classifier to the subcategory classifier decreased overall accuracy, it allowed some categories to be better classified. For instance, the category classifier was able to classify both religion and social science questions with an accuracy of 60.8%. The subcategory classifier, however, was able to classify religion questions with an accuracy of 85.4% and social science questions with an accuracy of 86.2%, even though the same questions and labels were being used. This improvement could be partly due to the subcategory model having a more balanced dataset, but its ability to better differentiate content within other categories and generate more specific criteria for belonging in a given bin increased the accuracy of classifying other questions.

When trying to classify the test data using cosine similarity, the accuracy was only 67.5%. This increased to 84.1% when considering both the largest and second largest cosine similarities. Even when the cosine similarity-utilizing model was allowed to consider the “top two” choices, the Naive Bayes classifier still outperformed it, showing that overall this was a more successful way of classifying the questions. The intuitive explanation for why this is the case is that the Naive Bayes model relies on many smaller vectors that can aid classification, rather than one large vector.

The subcategory classifier only performed at a 58.71% accuracy on the curated set of trivia questions, with the confusion matrix shown below.



Confusion matrix for the subcategory classifier on the trivia dataset.

The model does not perform terribly considering the large number of classes that exist, but it still is far from working at a good enough level to warrant any real-world use. There are two main reasons for this poor performance. The most obvious reason is that it was not trained on this type of corpus, and so the definitions of what belongs to a certain subcategory will differ. For example, the question “Born in London in 1478 what was the name of Henry VIII’s Lord Chancellor author of ‘Utopia’ who was executed for refusing to deny Papal authority? He was canonized in 1935.” should be categorized as European history according to both intuition and the classifier, although the trivia website labels it as a religion question. The other is that these trivia questions are noticeably shorter than tossups, with the majority of them being just one sentence. They therefore contain much less information for the model to work with, which makes classification more difficult.

**Conclusions**

Overall, models were successfully trained at the task of classifying questions both by category and subcategory, with the Naive Bayes approach performing better than relying on cosine similarity. These findings coincide with the visualizations of the data, which showed that the tossups are largely separable by category and subcategory. The most comprehensive model does not translate to other types of trivia questions as easily as predicted, but it could with some potential changes. Training on different corpora such as more trivia questions, or even other open-sourced academically focused datasets like Wikipedia could lead to more robust classifiers.

Certain aspects of the research process proved challenging. Although there exists thousands of questions for most categories, QuizDB’s generated json files are limited to only 750 tossups. Gaining access to these additional questions would strengthen the training dataset and vastly improve model performance. There does exist an API that allows access to all questions, but even after reaching out to the administrator it remained unusable at the time.

One other limitation of this study was the lack of subcategory labels on many questions, which limited the amount of subcategories that could be studied. For example, although a psychology subcategory does exist within the social science category, there are only 21 labeled psychology questions in QuizDB. Because this number is so small, there was no feasible way to incorporate these into the classifier in a balanced manner. If somebody were to manually label a few hundred tossups in these currently small subcategories, then they could be included in the model, leading to a classifier that can comprehensively categorize every type of question.

**References**

Iyyer, Mohit, et al. "A neural network for factoid question answering over paragraphs." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

Kuang, Raynor. QuizDB, PACE, [http://www.quizdb.org](http://www.quizdb.org/)/.

Osman, Addin, and Anwar Ali Yahya. "Classifications of exam questions using natural language syntatic features: A case study based on Bloom’s taxonomy." *Proc. 3rd Int. Arab Conf. Qual. Assurance Higher Educ.*. 2016.

Tan, Guan Kheng. *Web Based Application of Examination Question Classification*. Diss. UTAR, 2019.

*Trivia Bliss*, Amazon Services, 21 Oct. 2020, https://triviabliss.com/.