Machine Learning Improvements

Although the machine learning models gave a satisfactory performance last semester, the team thought the accuracy of the models were not enough to be utilized by the public. Another factor that made us improve the machine learning models was because the focus of this project is the study of machine learning and improving accuracy has always been the prominent stigma of the project. Therefore, we wanted to show our research on how one could improve an existing machine learning model.

Training a machine learning model is a skill that one could achieve by learning the engineering knowledge behind the model. Often, the data analyst will get the most out of a dataset through optimizing model parameters, superior feature extraction techniques, and smart data pre-processing approach. Our team found the best parameter values and feature extraction method for both datasets. While Model 1 got a satisfactory result, Model 2 did not get the accuracy we wanted. For the rest of this section, we will focus on how we are improving Model 2, followed with a discussion on Model 1, and then we will briefly discuss on how we can enhance the accuracy in the future.

The team used the 20newsgroup [1] dataset to train the current Model 2. Provided by scikit-learn [2], we used the linear support vector classification as the model. Using the LinearSVC library, we ran around 30 possible parameters before settling to the best result we got, which is the current version of the model. We are confident that the parameters and the vectorizer we chose were the optimum configurations for our model. Because of that, we can conclude that we would need more data in order to improve this model.

The 20newsgroup dataset has been around from 1995 [1]. The age of this dataset is a problem considering that we are dealing with up-to-date articles and sentences from people that live in 2020. We decided that the only way to improve our model is to train it with a newer, modern dataset. Since we previously mentioned that the configurations of the model are optimized, adding new dataset is the only option. The new dataset must be compatible with our current model and well informed with the current news.

There are two options to gather the data; by downloading a new dataset, or by gathering our dataset through a media or expertise. In this paper, we will give our report on how we attempt to go through the options and how we finally able to prove that we can increase the overall performance of the model.

Internet Dataset

Our existing models have specific inputs and can only train on a definitive dimension. We pre-process the dataset so that the model only trained on a list of sentences or objects in 20newsgroup, and the target was a series of numbers that represent the classes. If we downloaded a new dataset from the internet, we would need to find a dataset that focuses on natural language sentences and its categories. Unfortunately, we did not find many datasets that meet our criteria and requirement.

We found three datasets on Kaggle that potentially could be added to our dataset. The datasets were the Huffington Post News Category, India Headlines News, and the BBC News dataset. For each of these datasets, we pre-processed them accordingly so they would be trainable by our existing Model 2. We did not change any parameters of the model since we want to know on the impact of the new dataset.

Only the sentences and the categorical target were kept for the model training and testing. For datasets that have both article titles and description, we concatenate them into a column called a sentence. Then,

the code removed every column within the dataset except the sentence and the target (which is the categories). We used scikit-learn libraries to factorize the target and vectorize the sentences. The vectorized sentences were used for the model training purposes. Also, the train-test split was 20% for the test set for each of the dataset (and 80% for the training). Both the training and testing set from the downloaded dataset were concatenated to the 20newsgroup's training and testing datasets, respectively.

We slightly pre-processed the Huffington post dataset differently from the other because the classification given by the source was not reliable. Several classes encapsulate the same category. For instance, there were a class called ARTS & CULTURE and another class called CULTURE & ARTS. Because of this inconsistency, we decided to combine some of the classes before factorizing the target. These are the list of those classes:

- HEALTHY LIVING with WELLNESS
- QUEER VOICES with GROUPS VOICES
- BUSINESS with BUSINESS & FINANCES
- PARENTS with PARENTING
- BLACK VOICES with GROUPS VOICES
- THE WORLDPOST with WORLD NEWS
- STYLE with STLE & BEAUTY
- GREEN with ENVIRONMENT
- TASTE with FOOD & DRINK
- WORLDPOST with WORLD NEWS
- SCIENCE with SCIENCE & TECH
- TECH with SCIENCE & TECH
- MONEY with BUSINESS & FINANCES
- ARTS with ARTS & CULTURE
- COLLEGE with EDUCATION
- LATINO VOICES with GROUPS VOICES
- CULTURE & ARTS with ARTS & CULTURE
- FIFTY with MISCELLANEOUS
- GOOD NEWS with MISCELLANEOUS

After factorizing, vectorizing, and splitting the Huffington post dataset, we concatenated the dataset with the 20newsgroup. Then, we trained the model, and the full result can be observed in model2_log_huffington.txt text file.

As the logs suggested, the performance of the machine learning model dropped significantly with 65% overall accuracy. Some of the new classes within Huffington post dataset got an f1-score of 0. Even when we trained several more linear support-vector-classification with different parameters, the accuracy of 65% seemed to be the best accuracy we can get.

We were not satisfied with this result, so we consulted Kaggle notebooks [3] to look for further improvement. However, it seems that even the best accuracy is at 65%. We decided that the dataset did not have the best quality for NLP machine learning.

We moved on to the India headlines news dataset. As before, we always start with data analysis when training a model. After we removed 6% of the dataset that has the class 'unknown', we are sceptical on the statistics of this dataset. For starter, there are 1015 unique classes within this dataset. Upon closer inspection, a majority of the news articles were local Indian articles. Since we are building an Australian-based English system, we decided to ditch this dataset for its exclusivity and the low quality.

We moved on to the BBC News articles. The dataset has five new classes, which in theory should be more robust than the previous datasets. We pre-process the dataset with minimal alterations, as we only combined class "politics" with "talk.politics.misc" in the 20newsgroup. For the dataset, we did a 66-33 split for the training and testing set, since this split seems to give the best possible result. The results for the model are stored in mode2_log_bbc.txt text file.

Based on the results provided in the text file, the overall accuracy is 82%. The number suggests there is no significant improvement from the old Model 2 in terms of overall accuracy. Upon closer inspection, we found out that the model has better understandings when classifying the new classes. We emphasized on these f1-scores:

- Business: 0.95
- Tech: 0.92
- Sport: 1.00
- Entertainment: 0.92

Moreover, the class "talk.politics.misc" has experienced an increase as well. In the old Model 2, the class has an f1-score of 0.67, while the new Model 2 with BBC news the f1-score is 0.71. Because of this, we could conclude that the BBC dataset has a high potential as the model improvement in our system.

Dataset Gathering by the System

Aside from the internet, we also sourced our dataset. Our website has a feature where the user will be prompted to fill in some information before we give out the recommendation. This information is recorded and stored so that the machine learning models can train from the user's inputs.

The philosophy of the system is by training the machine learning models with a little additional data at a time; the model will gradually improve as the website collects more data from the users. The drawback of collecting our dataset is the limitation of resources, such as time and medium. We only have a limited time to make our dataset, and usually, the dataset collected in a small amount of time will not be as many as the dataset gathered from the internet. That is why it is essential to consider this process as a gradual process instead of immediate improvement.

Since getting public data would be very hard at the current stage, we filled in our sentences to the system. We made a dataset that contains 400 data points, and each class in 20newsgroup got 20 instances within the dataset. We constructed the sentences from the RSSFeed collected in 25/09/2020. In order to get real sentences, we did not copy and paste the sentences, but rather rephrase the article titles. Also, some of the sentences came from our thoughts and reasoning. The new dataset is stored in newdataset.csv CSV file.

Since the new dataset only has 400 data points, we put all data points of the new dataset into the training set. The pre-processing, feature extraction and training process were straightforward. The resultant model is stored in model2_logs_newdataset.txt text file. As expected, the accuracy of the model stayed

the same, which is 82% of the overall accuracy. The data from 20newsgroup vastly outnumbered the new dataset.

Testing the Improvement

The overall accuracy presented in the logs did not suggest that there is an improvement for both our dataset because the best results for both approaches are at 82% overall accuracy. The number stays the same as the old Model 2. However, the reason why the overall accuracy stayed the same is that the data from 20newsgroup dominate data points within the testing set. Statistically analyzed, the accuracy presented in the logs are not representative of the overall improvement because the majority of the datapoints are from 20newsgroup.

There are a few disadvantages when relying on 20newgroup dataset. The most obvious argument is the age of the dataset, where 20newsgroup first recorded appearance was in 1995. Our system updates the article database regularly. Consequently, the system would need updated information on the current global state. Countless new words and trends appeared between 1995 and 2020, and arguably a dataset that was created in 1995 will not be able to train a model that comprehend new dictionaries and universal progressions. In short, the machine learning models will have a troublesome time when predicting current article titles, and the accuracy will not be maximized.

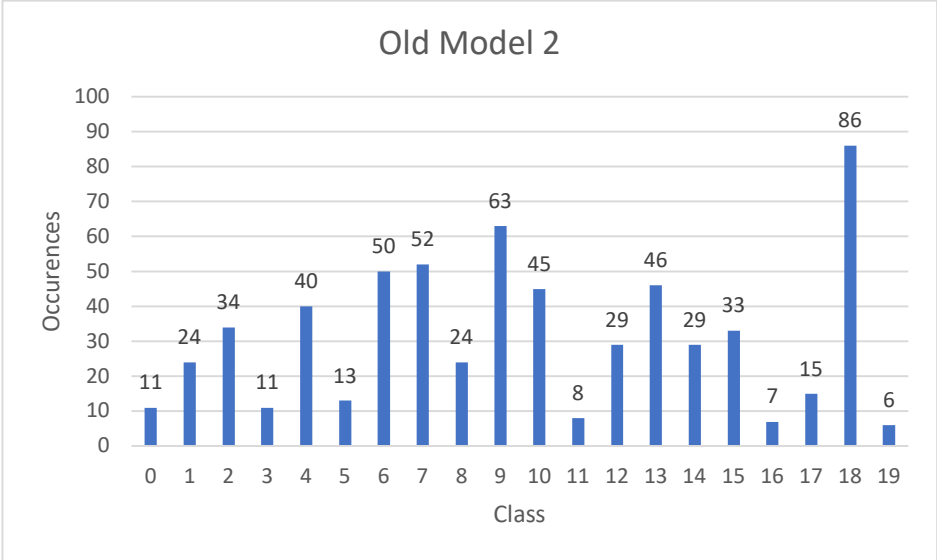
Aside from the outdated consideration of the 20newsgroup dataset, the numbers represented in the logs did not represent the overall improvement of the models. Most of the test set came from 20newsgroup, so the statistics will be biased for the results that came from 20newsgroup dataset. The dataset, 20newsgroup, has a distinct format for its content, therefore relying on the results presented in the logs will not be good.

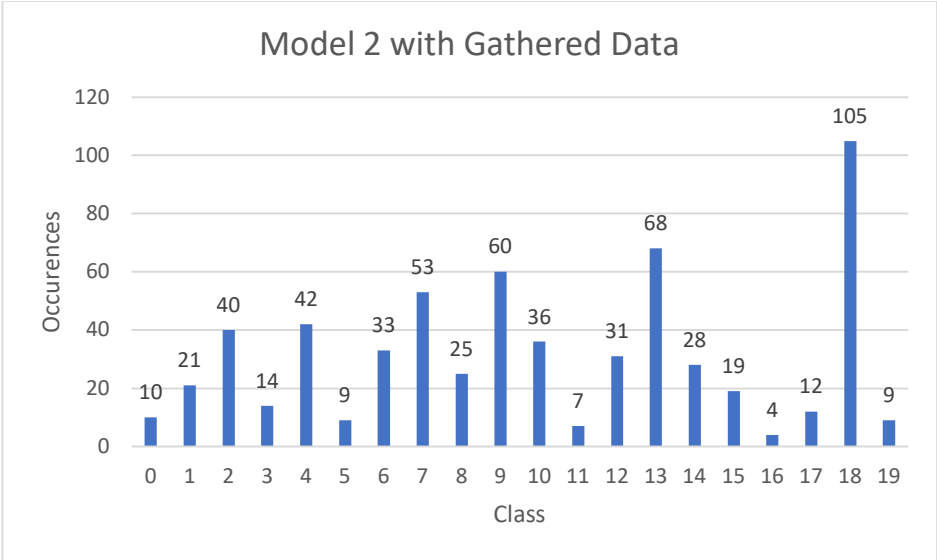
Because of the reasons presented above, the team construct a solution to test the real improvement of the machine learning model. The test must be from an outside source, where it comes from neutral ground. All datasets, which are 20newsgroup, the BBC news dataset, and our dataset, must have no connection to the test data. The best source for the test data is the RSS feed.

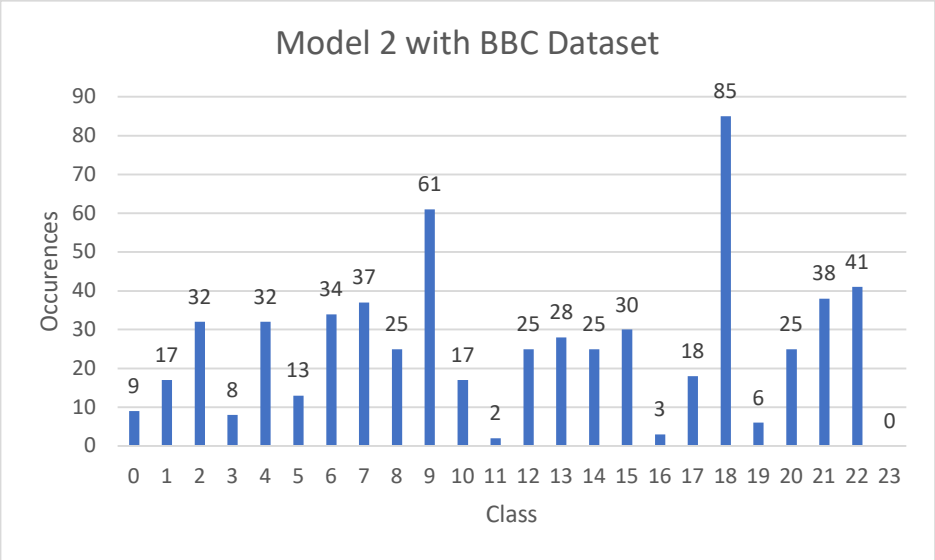
There are a few advantages to test the performance of the models from article titles and descriptions that are collected from the system's RSS feed. Firstly, the machine learning models are predicting the article's classification within the system. Testing the dataset with the article would be a simulation and audit for the machine learning models since the developer can record the real accuracy and performance on the model. Secondly, the titles and the descriptions of the articles are stored in a database that has not been used for model training. The situation provides a neutral ground on testing the data.

The process started with updating the article-database, so the system refreshed the database with fresh articles that are current. All models will run and predict each of the articles; then the results are stored in a CSV file. The team performed a data analysis on Excel based on the results of the differential between the old Model 2 and the new Model 2.

For this testing, we did not include the BBC news dataset. The reason why we did not include the BBC news dataset is that hat in there are two problematic classes, "Tech" and "Sport". Both classes are broad in their respective fields. Meanwhile, 20newsgroup has nine classes that can be sub-classes for "Tech", and two different classes that are sub-classes of "Sport". When inspected, some classes are under-represented when we include the BBC news dataset. Here is an RSS feed that has taken in 18/10/2020:


Old Model 2
Occurences
Class
100
90
80
70
60
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
11 24 34 11 40 13 50 52 24 63 45 8 29 46 29 33 7 15 86 6


Model 2 with Gathered Data
Occurences
Class
120
100
80
60
40
20
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
10 21 40 14 42 9 33 53 25 60 36 7 31 68 28 19 4 12 105 9



As we observed, the occurrences for class “Tech” and “Sports” found themselves as classes with high occurrences relative to the other classes. From the graphs above, we can see that classes 1, 2, 3, 4, 11, 12, 13, and 14 are having fewer occurrences in BBC model than other two models, and some classes have detrimental declines. The trend happened because all of those classes could be defined as class 21, which is the “Tech” class. Technically, each article that belongs to class 1, 2, 3, 4, 5, 11, 12, 13, and 14 are also class 21. Comparing the models will not make any sense.

Because of the justification presented above, we compare only the old Model 2 and Model 2 with our dataset. The results are presented in the data_analysis_differential.xlsx Excel file. In the document, the team made a new sheet that contains all articles where the two models give different classification. We analyze the list of these articles and look at each article titles and descriptions in that list.

The team read each article manually and score every classification from both models. For instance, an article that discusses a new law in the lens of an atheist could be classified as class 0 and 18. The scoring method was simple; 1 for correct classification and 0 for incorrect classification. Because there were ambiguities when it came to particular articles, the team devises some guidelines for scoring the articles:

- All COVID-19 related articles are in class 13.
- All sports related article that are not either baseball or hockey could be specified as class 9 and class 10.
- All science related article that are not either cryptography, electronic, medicine, or space could be specified as class 11, 12, 13, 14.
- All religion related article that are not Christian are considered class 19.

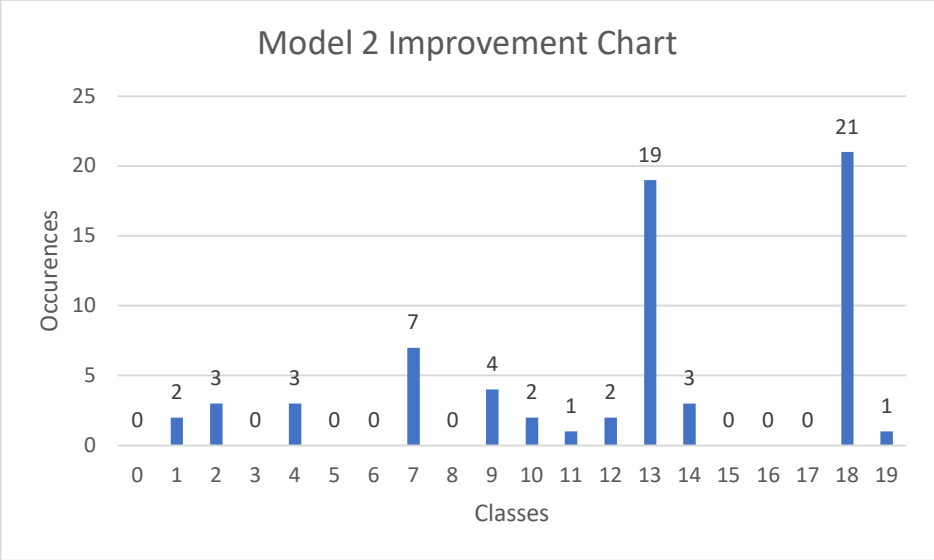
Although the team had the guidelines, some articles and classification remain ambiguous. We enforced the “benefit of the doubt” policy, where if classification is not “clear” correct or incorrect, we will give it a score of 1.

There were 626 observed articles, where 118 of them has different classifications by the old Model 2 and the new Model 2. Among those 118 articles, the team found that the old model successfully identified 20 articles, whereas the new model got 82 classifications correct. Upon closer inspection, there are six

different instances where the old model got the correct classification, and the new model got it wrong. Nevertheless, the team found the number of improvement (old model got the classification wrong while the new model got classified it correctly) is 69 out of 118. If we want to know the overall improvement, the number would be 11.02%, where:

$$Overall\ Improvement = \frac{The\ number\ of\ improvement}{Total\ articles}$$

The spread of the improvements is presented below:



The articles were gathered amid global pandemic and political election (October 2020). Logically, articles about Coronavirus and politics are more frequent than any other news article. In the graph, the new Model 2 has demonstrated an updated knowledge of medical science and political stances than the old Model 2.

Because of this, we can conclude that statistically, our additional dataset was able to rise the accuracy of the model by 11%.