Implementing More Efficient Models to Reduce Knowledge Obsolescence in Social Media

Enterprise Computing Project Final Report, Summer 2024

Dongjae Lee

Spencer Starks

Table of Contents

Introduction

Over the past few decades, the internet has gone through explosive growth, providing access to far more people every year worldwide. This surge in connectivity has revolutionized global communication, enabling people from different regions to interact and share information at unbelievable speed. However, these reduction in barriers also brings significant challenges, particularly the rapid spread of misinformation, including outdated or incorrect information. Social media platforms, driven by advanced algorithms, play a pivotal role in controlling information by classifying and recommending content to their users.

For machine learning models to maintain optimal performance, they must constantly adapt to the dynamic data they process. This requirement is complicated by a phenomenon known as Knowledge Obsolescence (KO), where the relevance and accuracy of the data used to train these models diminish over time. KO presents a critical challenge, as the underlying data becomes obsolete, the models' ability to make accurate predictions and recommendations degrades, allowing for greater spread of unreliable information.

Throughout the course of this project, our team aimed to produce models that can create datasets that reduce KO without human intervention, and we exceeded our expectations. Through the application of these methodologies with a robust testing and visualization network, we have compiled a comprehensive realization of our achievements. Throughout this final report, we will cover the various intricacies of our fully completed project as well as the skills that we learned along the way.

Motivation

Addressing the issue of KO is critical for keeping the integrity and reliability of machine learning models, particularly in the context of social media and information based platforms. This project focused on tackling the challenge of Knowledge Obsolescence through the use of weak supervision techniques for COVID-19 fake news. Weak supervision leverages a range of heuristic and semi-supervised methods to label data, reducing the dependency on large,

manually curated datasets. Through the use of weak supervision algorithms, we can improve the process of identifying outdated information and automatically annotate outdated information which is much faster than manually going through each dataset. This project serves to simplify the process of updating outdated information to make sure that users are exposed to less misinformation while browsing online.

This form of implementation is essential due to the rampant spread of misinformation across the internet. Many major societal topics have become astroturfed by bots designed to spread misinformation and people not properly educated in the topic resulting in ignorance dominating the information era. Many people are confused with basic facts with a poll by the Pew Research Center showing about 88% of US adults have stated that social media caused at least some confusion about basic facts of current events.

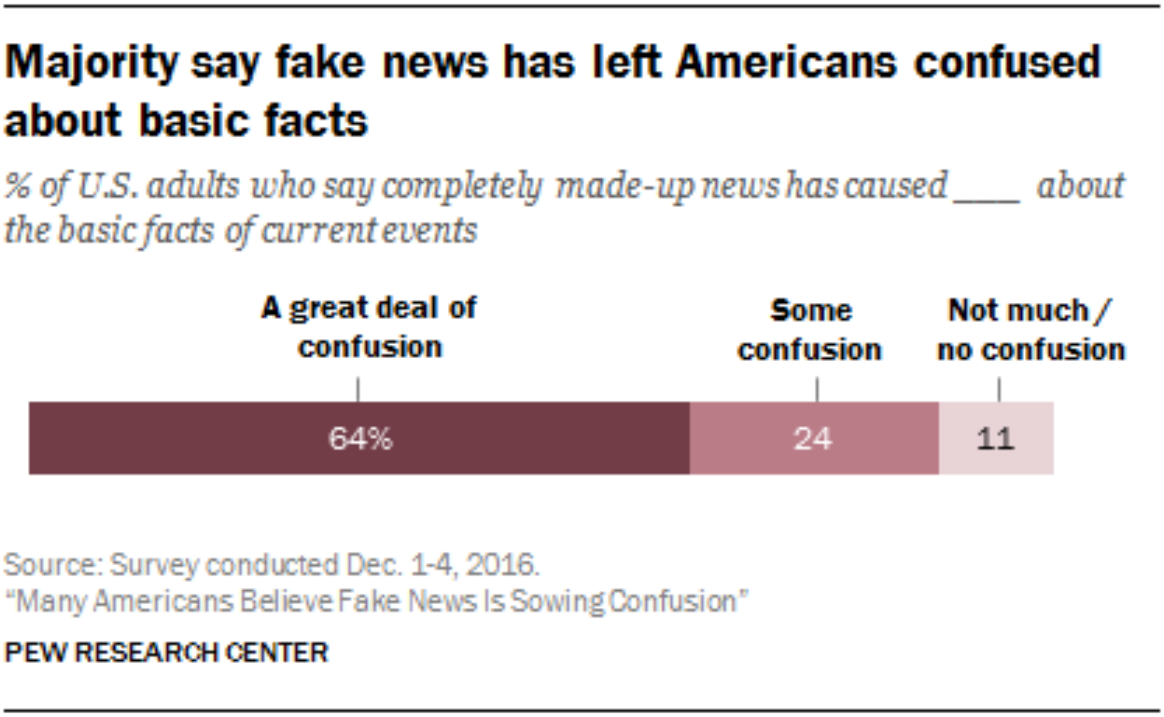


Figure 1: Amount Americans confused by misinformation (Barthel et. al 2016)

With massive events such as the United States presidential election on the horizon, a fast and reactive social media filter is essential in preventing fake news. Our hopes of a model that adapts quickly and is not hampered by knowledge obsolescence can quickly understand bias and fact from fiction when

analyzing social media content. This would in turn allow for the model to reduce misinformation from users' news feeds to promote factual information and disregard fake news.

We propose incorporating a weak supervision model to detect and mitigate the impact of outdated information in real-time to improve the adaptability and resilience of ML models without the need for many resource-intensive retraining sessions. This approach will reduce the need to retrain models to continuously stay up to date to the current news cycle and remove misinformation.

Related Work

There has been multiple pieces of work related to our project both in the domain of machine learning and real world applications based on the Snorkel framework. Snorkel created a revolutionary new angle to machine learning where datasets no longer need to be entirely labeled by hand, and with the use of weak supervision, machine learning models based on Snorkel can now label relevant data to be used to train future models. Many cases have applied these techniques to a staggering degree.

Farchi et al. (2021) presented a method to correct model errors in chaotic systems, such as weather forecasting, where the simulation models need frequent updates due to the dynamic nature of weather patterns. Their approach involves using machine learning to assimilate new data and adjust the model's predictions, which fell in line with our goal of diminishing knowledge obsolescence by continuously updating models with new information.

Ratner et al. (2017) developed the Snorkel framework, which as mentioned earlier, is a weak supervision framework that significantly reduces the time and effort required to build and update machine learning models. By leveraging weak supervision, Snorkel allows for rapid creation of training data, resulting in models that are both faster to build and more accurate. This research was backed by quantitative data showing that researchers were able to “build models 2.8× faster and increase predictive performance an average 45.5% versus seven hours of hand labeling”. This framework has been successfully applied in various areas, including text extraction and chemical reaction identification, proving both its

versatility and effectiveness in handling large-scale data (Ratner et al., 2017; Mallory et al., 2020).

Furthermore, Snorkel could be applied in more industrial settings, as documented by Bach et al. (2019), showing that weak supervision can improve both the scalability and efficiency of machine learning tools. This case study demonstrates the practical benefits of integrating weak supervision techniques in a real world scenario, which would be relevant to our project as we focus on reducing misinformation on social media platforms.

These studies show the effectiveness of using weak supervision to improve model accuracy and relevancy, thus it can tackle our project's challenge of reducing knowledge obsolescence. Our research builds on these insights to develop robust models that remain current and effective in dynamic environments such as social media.

This has led to the large adoption of Snorkel across the industry as well, with many major corporations picking up the framework. According to WirelessNews (2023), Microsoft has implemented Snorkel natively within their Azure cloud computing system to improve use cases for data analysts and machine learning developers. With many more companies picking up the tools and making Snorkel a mainstream network, the research and projects developed in relation to our own grows.

Underlying Design

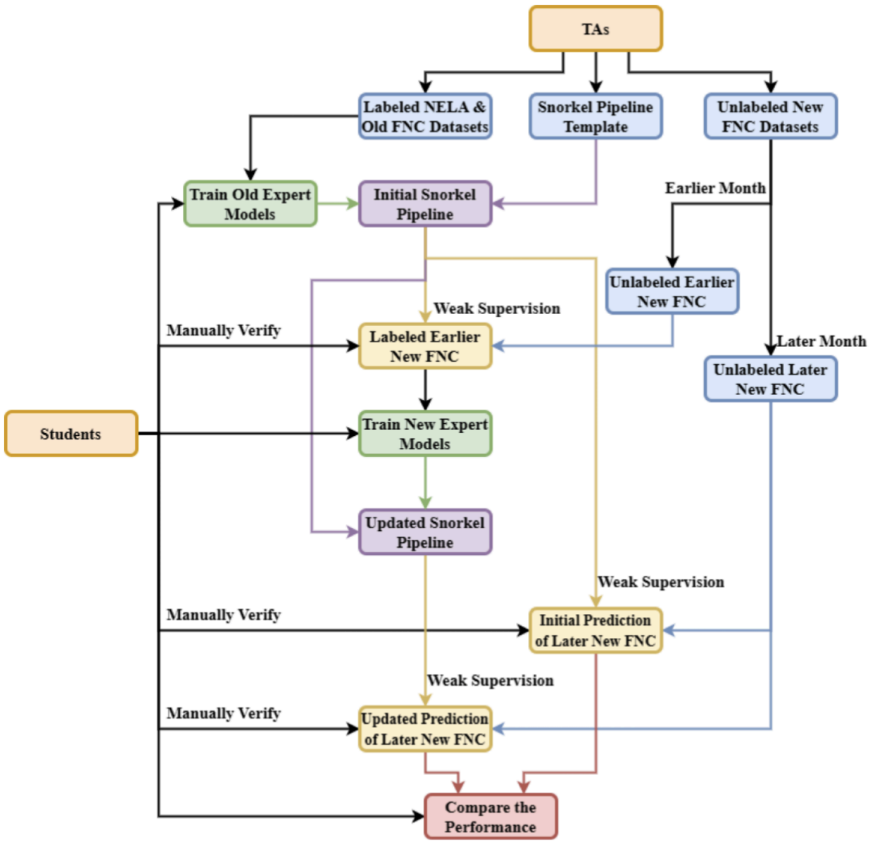


Figure 2: Model Training Methodology and Workflow

The underlying design patterns came to fruition due to many iterations of different approaches to model design and effectiveness. This model methodology and workflow was achieved through multiple stages of our project as we slowly built up the project from the beginning to the end of our workflow. We began by determining and training each model to determine how effective they are at learning and accuracy in comparison to the models.

We additionally applied these models to a weak supervision architecture known as Snorkel. With the most effective models working in tandem with Snorkel, we can produce labeling functions that produce new datasets that are effectively the same as hand labeled datasets. This would achieve the reduction of human made datasets and allow for our model to produce new datasets that

are removed from the problem of Knowledge Obsolescence and produced at a much faster rate. We can begin by looking at the information collected on the various datasets and models we worked with.

Snorkel Pipeline

The snorkel pipeline serves as the core of the project and is provided by the TAs. This pipeline contains pretrained models from the NELA dataset and is the baseline for further improvements by other labeling models. The models are able to label untrained data and has the logic for data preprocessing and labeling and is the baseline for comparisons with more updated models. Throughout the course of his project, we could apply multiple different model combinations to the pipeline to quantify how the quality of the produced datasets would be.

Dataset Labeling

Before the implementation of the labeling functions, we first develop our testing variables that serve as our ground truths. These datasets are json files that contain information regarding tweets during a certain period of time and are tweets that were identified during the COVID-19 Pandemic. The datasets themselves are elaborated much more in depth further below in the datasets section of this report.

Labeling Functions

Throughout the project the development of labeling functions serves as the key step of improving the initial models that have already been provided. The labeling function serves to classify text based on what the model is able to discern from the text that it receives. Our project worked on the development of the GPT and GPT2 models by using the pretrained model of GPT2 to develop a more robust labeling function that is able to use more parameters within the GPT model.

Pipeline integration

Once our labeling functions have been implemented, we would integrate our model with other high quality models such as ALBERT, and BERT which we also train to compare with our base model. Once the pipeline has been

integrated, we can compare the results of the updated pipeline with the original pipeline to check if there are any improvements with our updated Snorkel Pipeline.

Updated Model Comparison

Once our Snorkel Pipeline has been updated, we would compare the results of our updated models with the original labeling functions and see if the changes we made resulted in an improvement of the model. The comparisons would measure the effectiveness of our work throughout the semester by measuring the overall accuracy of the system.

Datasets

There were three datasets used throughout the entirety of the project. Consisting of the NELA, Old FNC, and New FNC datasets, each set provided a distinct lens for the models produced and trained throughout the course of this project. Each dataset was collected in a distinct manner providing different insights in both the collection methodologies and time of data being collected providing a robust amount of information to train the models on.

NELA Dataset

The NELA, News Landscape, dataset is a dataset consisting of tens of thousands of datasets from news networks of all kinds. This dataset is labeled to analyze whether the news article would be considered reliable. This in turn provides confirmation on whether the article is more likely to be misinformation or a source that is trustworthy. With this information, the datasets can effectively be used in tandem with this project for COVID-19 related misinformation and whether its related news can be considered misinformation. This dataset provided an excellent baseline for the research conducted on this project. For our project in particular, we worked with subsampled portions of the NELA dataset based between November and December of 2020 providing insight into the latter half of the first year of the COVID-19 pandemic.

Old FNC Dataset

The Old FNC, Old Fake News Challenge, dataset consisted of a subsampled set of 500 text entries consisting of tweets labeled either as 1 for being trustworthy information or 0 to represent misinformation. This dataset provides a much stronger correlation to the goal of identifying misinformation related to COVID-19 as they are more focused on the subject at hand. The datasets we worked with ranged between April and May of 2021 providing context in the second year of the COVID-19 pandemic. Overall, this dataset provided a strong baseline for how we would approach the newly created FNC Dataset to provide an additional temporal context to the subject of this project.

New FNC Dataset

The New FNC, New Fake News Challenge, dataset was hand labeled and produced by our team focusing entirely on misinformation related to the COVID-19 pandemic. Our dataset was based on the months of October and November of 2021 focusing on the end of the end of the second year of the COVID-19 pandemic. We had a massive dataset consisting of hundreds of thousands of entries that required heavy cleaning and analysis to properly label the data. We began by cleaning the data with a preprocessing approach that is further elaborated upon in the next section. Once the data was cleaned we could then manually go through the data and label information as 1 if the data was reliable or 0 if the data was misinformation.

Dataset Annotation

Data Preprocessing

We noticed that there was a large number of retweeted data labeled with "@RT" at the start of each tweet. In order to remove any redundant posts, we removed them through a python script we created. This dramatically reduced the number of posts that we could parse through on the json file. In addition, we also removed all non-English tweets with an additional python script applying effective third party libraries that helped identify the language of the text. This dramatically helped with the data labeling in the future as it allowed for us to only label posts in English which also helps in removing any potential variables in our

experiments. In addition, we converted our json file to a csv file with 2 columns, the text file and the label file. This substantially reduced the number of data that was shown from the dozens of parameters stored on the json file such as the user name and time posted into just the text in the tweet which then reduced the file size by a lot. The empty label column made it easier for us to label the resulting filtered FNC data.

October 2021 Annotation and Analysis

Number of real news annotations: 339

Number of fake news annotations: 345

Tweets	Label
Because they want to be part of the 1/10th of the population that is not depopulated per the plan of taking out 9/10ths of the pop. in this eug. agenda. And they are always the privileged pointing 1 finger but pointing 5 fingers back it seems. https://t.co/sBByIuVuCc	0
Quick: get the horse paste and blue light probe. https://t.co/zAP9N93sMN	0
Shame he's vaccinated. https://t.co/hZGCLVcx4a	0
@mediahunter 1,143 new local Covid cases and 3 new deaths in Victoria. Down from 1,438 new cases and 5 new deaths yesterday.	1
@NDarago_stvm59 1. That’s a massive broad brush. For 17 year olds, the death likelihood is 1%. For 50-64 year olds it’s 10%. For 65-74 year olds, it’s 12%. + covid can have long lasting effects on anyone, which can be avoided if everyone vaccinated	1
Many psychologists have attributed pandemic brain fog to chronic stress’s impact on the prefrontal cortex, where it can impair concentration and working memory. https://t.co/mH2jTHaUNW	1

November 2021 Annotation and Analysis

Number of real news annotations: 287

Number of fake news annotations: 213

Tweets	Label
@NickAdamsinUSA There is no such thing as a Covid vaccine.	0
@seanhannity So she has the flu so what that's all COVID is just a different type of flu	0
@EROTHCJ5 I don't wish Covid on anyone but it is obvious that a vaccine mandate isn't the answer. Therapeutics properly prescribed and getting on with life is the only answer. https://t.co/X5cjcBe0DP	0
The number of new Covid-19 cases in much of the United States is declining, but Dr. Anthony Fauci says that children should still get vaccinated as soon as they are able. https://t.co/7qd4Ky0573	1
Sotrovimab is one of the new COVID-19 treatments that have been approved in Australia. Join @Austin_Health on Thursday, 4 November at 7 PM to hear from experts who has significant experience in managing and treating #COVID19 patients. Register here: https://t.co/OTBABgWWCO https://t.co/iZwEFljbls	1
RSPCA sees 166% increase in cat adoptions during the pandemic Read it here: https://t.co/HLtg5DFhfh https://t.co/VRzvPZbCSH	1

Throughout our manual labeling process with the large datasets we discovered many interesting insights. There was an insurmountable number of entries that were either retweets or written in a foreign language. We understood that parsing through hundreds of thousands of data points would be

overwhelming when we need to identify 1,000+ high quality pieces of data to train and test with. This resulted in our use of a retweet and english filter reducing the raw data from 650,000 entries down to 150,000 entries. This resulted in a 77% reduction of noisy and useless data significantly enhancing our progress for the first checkpoint of the project.

As seen in the samples above, we parsed through a multitude of different data points with factual information, labeled with 1, typically originating from news sources or individuals stating factual information. Misinformation, labeled with 0, were typically replies or statements from individuals who were not properly informed about the COVID-19 vaccine or conspiracy theorists spreading information that is not scientifically backed. Due to the structure of misinformation, most news sources would not be labeled as such as they need to remain accountable. Therefore, most misinformation was produced by individual messages or replies.

Models

Our team worked with various different models during this project. We worked with the ALBERT, BERT, GPT, GPT2, and RoBERTa models through various tests and analysis. Unfortunately, the RoBERTa model failed due to token errors preventing us from running the model as intended and we had to divert focus to the other mentioned models. Throughout the project we discovered various noteworthy aspects of every model we tested.

ALBERT Model

The ALBERT Model was produced off of every subsampled dataset and resulted with very mediocre training accuracies. The tradeoff between less computing power and the overall accuracy was too high as although the model ran much more quickly, the results that it gave out were invalid as it either predicted the text to be all true news or all fake news with small changes to the hyperparameters of the configuration file. Disappointed with the performance, we decided to distance ourselves from this model and focus on more successful models that produced quality training and testing accuracies.

BERT Model

The BERT Model was much better in terms of generating readable results as it was able to successfully make predictions on the testing dataset. The testing accuracies seem to max out at around 70% before succumbing to overfitting which was much better than ALBERT's singular predictions. We decided to leverage the BERT models even further with the Snorkel framework as we were happy with the generally positive accuracies on the datasets we were testing.

GPT Model

The GPT Model resulted in a similar outcome to the ALBERT model. While the training was rather fast compared to the BERT model, the training's accuracy would hit a standstill of 78%. We came to the conclusion based on our prior hypothesis with ALBERT that the model was not properly making decisions based on the provided dataset. After changing the hyperparameters of the GPT configuration, the training accuracy remained the same. The accuracy for this model sat around 60% which was below what we hoped for this model.

GPT2 Model

The GPT2 Model did an excellent job in producing positive testing accuracies without having to deal with overfitting. During the training process, the GPT2 model would reach a training accuracy of roughly 80% while the BERT model would typically reach a training accuracy of 95%+ in comparison. The BERT model would typically overfit in the latter half of the training while GPT2 kept a more conservative accuracy resulting in the final, the 10th epoch, GPT2 accuracy sit around 70% in accuracy while usually the 5th epoch of the BERT model would reach the 70% accuracy. Additionally, the GPT2 model would have a faster training time than the BERT model, however GPT2 takes up far more memory. Every BERT checkpoint sits at around 100 megabytes while a GPT2 checkpoint sits at around 1 gigabyte. Thankfully GPT2 was not plagued by the same training fatigue that occurred with the ALBERT and GPT models.

Model Configuration and Training

```
albert_config = {
   "data_param": {
       "dataset": "time_sorted",
       "max_data_size": -1,
       "batch_size": 32,
       "data_root": "/content/drive/My Drive/KO/",
       "train_datapath": "11-2020",
       "val_datapath": "",
       "test_datapath": "12-2020",
       "num_classes": 2,
       "filter_long_text": True
   },
   "model": "albert",
   "tokenizer": "albert-base-v2",
   "model_param": {
       "vocab_size": 30000,
       "embedding_size": 128,
       "hidden_size": 768,
       "num_hidden_layers": 12,
       "num_hidden_groups": 1,
       "num_attention_heads": 12,
       "intermediate_size": 3072,
       "inner_group_num": 1,
       "hidden_act": "gelu",
       "hidden_dropout_prob": 0,
       "attention_probs_dropout_prob": 0,
       "max_position_embeddings": 512,
       "type_vocab_size": 2,
       "initializer_range": 0.02,
       "layer_norm_eps": 1.0e-12,
       "classifier_dropout_prob": 0.1
   },
   "trainer_param": {
       "epochs": 10,
       "val_epochs": 1,
       "loss_func": "cross_entropy",
       "metric": "acc",
       "optimizer": "AdamW",
       "optimizer_param": {
           "lr": 5.0e-5,
           "eps": 1.0e-6,
           "weight_decay": 0.0005
       }
   }
}
```

The models that we had created were very similar in types of configurations that existed such as the data parameters that specified the training data and testing data, model parameters which defined the hidden layers of the model, as well as the training parameters which set the settings for our optimizer for the binary classification. As mentioned with our ALBERT Model subsection, during our initial model training and configuration, we used the ALBERT model to test our datasets. However the results that we got were not performing well so we tinkered with the model parameters, in particular, the batch size, the number of epochs, the learning rate and the weight_decay. Changing these values ultimately did not change the outputs of the ALBERT model and after discussing with the TAs we discovered that there was an issue with the model itself and ultimately moved on to other models.

On the other models, it was the learning rate and the number of epochs that were instrumental with getting our model to show better results overall. One prominent issue that we had with our models was that the models were prone to overfitting as the training accuracy of the models would reach 95%+ by around 7 epochs and then overfit afterwards. To mitigate the effects of overfitting we tried to limit the number of epochs to stop the model before it overfitted and by reducing the learning rate so that the model would train more slowly and reach its optimal state much more gradually. These changes were overall very similar to each other and the results of our model as well as the visualizations that we incorporated are further discussed in the next section below.

Model Results

The training and modification of the models to a reasonable extent allowed us to test out our models on the New FNC datasets that we had labeled earlier in the project. We included relevant records related to the ALBERT Model to provide more insight into the degradation of the model.

Training Records of Initial ALBERT Model

NELA 11-2020 ALBERT Model

Training Epochs

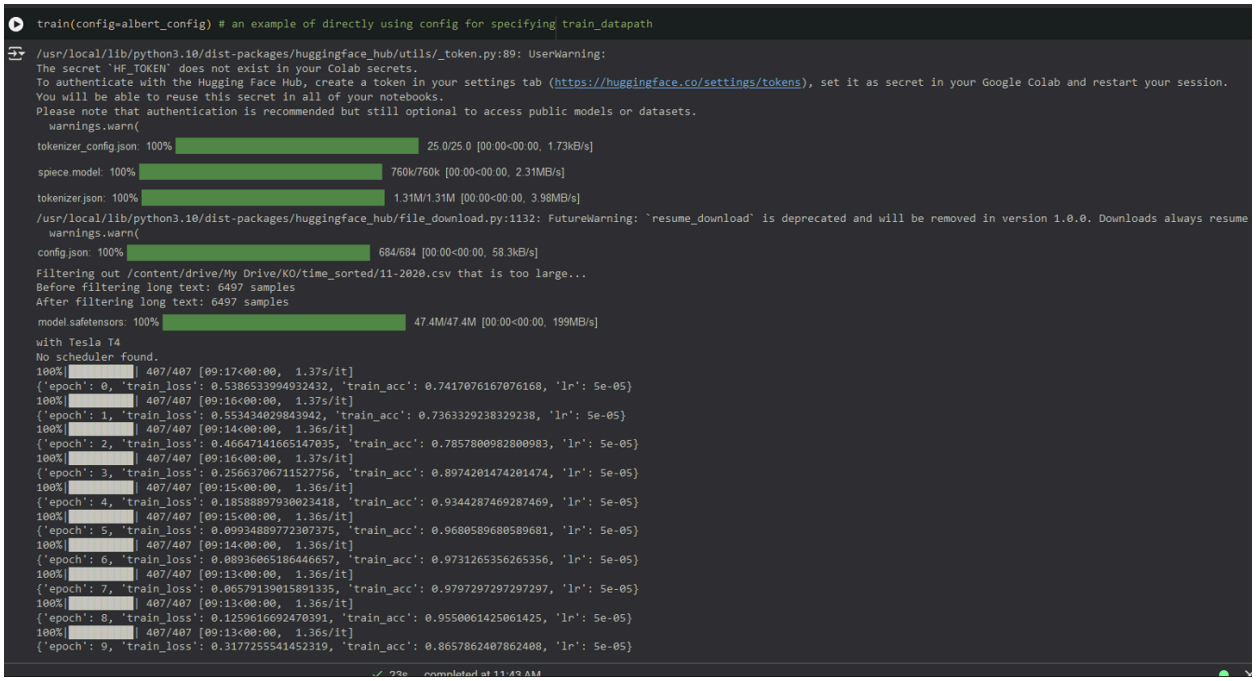


Figure 3: Training Epochs of ALBERT Mode based on NELA 11-2020 dataset

Testing Results on 2021-10-01 new FNC dataset

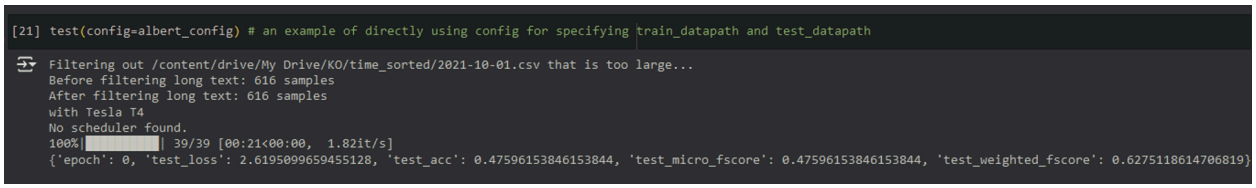


Figure 4: Testing Results of NELA 11-2020 ALBERT Model against the 10-2021 New FNC dataset

The prior figures represent the typical process training and testing the various models throughout the course of our project. Regarding figure 3, during training, multiple epochs of the model will be run retraining multiple times during each singular train. After each epoch, the training accuracy would be displayed representing how accurate the model is on the dataset that it is training with. The training loss gives details on whether there is potential overfitting or underfitting

allowing the training to determine if the model is going to have a high quality testing accuracy.

The testing results in relation to figure 4 provides insight into the metrics that are produced when looking at a fully trained model. The training accuracy shows how well the model can label pieces of data in a blank dataset and the labeled version of the dataset compares against it showing the accuracy in a percentage value. The rest of the information given by the tests were not useful in the case of this project, but are useful metrics for other forms of machine learning.

NELA Training and Testing Accuracies Based on ALBERT Model

ALBERT Model	NELA 11-2020	NELA 12-2020
Training Accuracy	97.3%	65.6%
New FNC 10-2021 Testing Accuracy	47.5%	47.1%
New FNC 11-2021 Testing Accuracy	57.2%	57.5%

Old FNC Training and Testing Accuracies Based on ALBERT Model

ALBERT Model	Old FNC 4-2021	Old FNC 5-2021
Training Accuracy	70.3%	75.7%
New FNC 10-2021 Testing Accuracy	49.1%	48.1%
New FNC 11-2021 Testing Accuracy	56.8%	57.5%

The information above provided significant insight into the issues we came across with the ALBERT model. Through debugging, we eventually came to the

conclusion that the model overall had poor results. We decided to focus on researching BERT models instead and visualize their accuracies below.

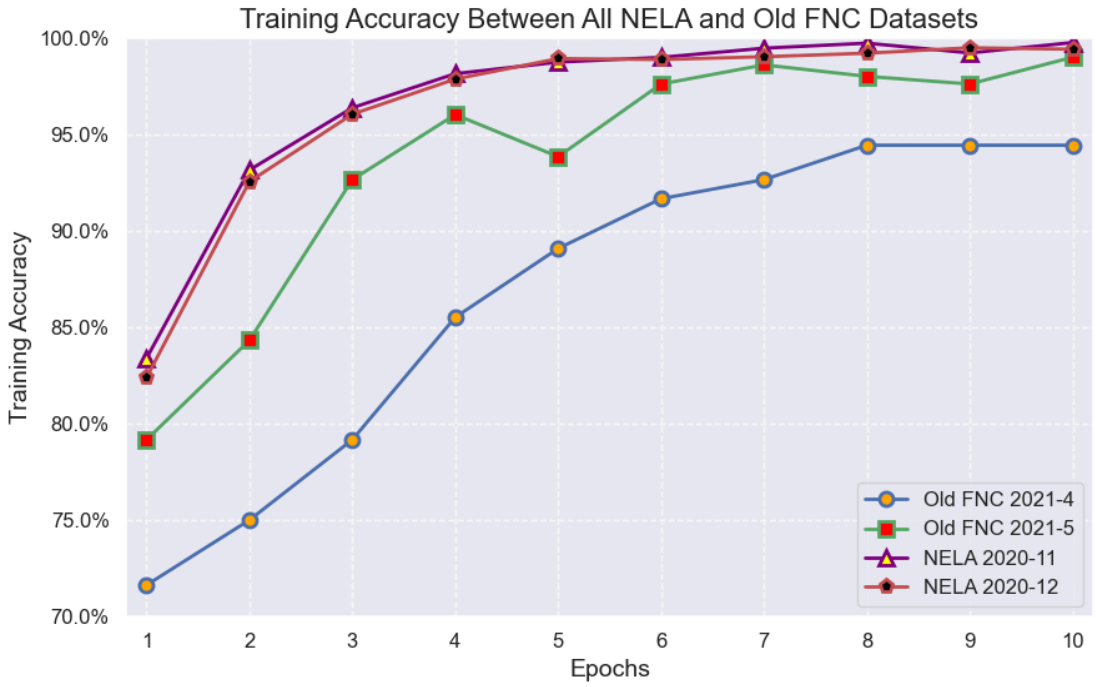


Figure 5: BERT Training Accuracy During Each Epoch

This graph showcases the increase in training accuracies of every dataset over sequential epochs. The training accuracies would typically begin with a low point and slowly build up to a very high accuracy. The Old FNC datasets did a good job not immediately jumping to a near perfect accuracy as that depicts their models must have been properly trained. The most important note is that both NELA datasets rapidly grew in accuracies over fewer epochs showing a potential of overfitting with their data.

We were able to hypothesize potential reasons for why the overfitting occurred and the most logical conclusion must have been due to the poor references of data of the NELA datasets. They were less connected to a specific topic and more spread apart in ideas on what should and should not be labeled. This must have resulted in the models not fully understanding what should and should not be collected and instead focus on selecting the same entries repeatedly.

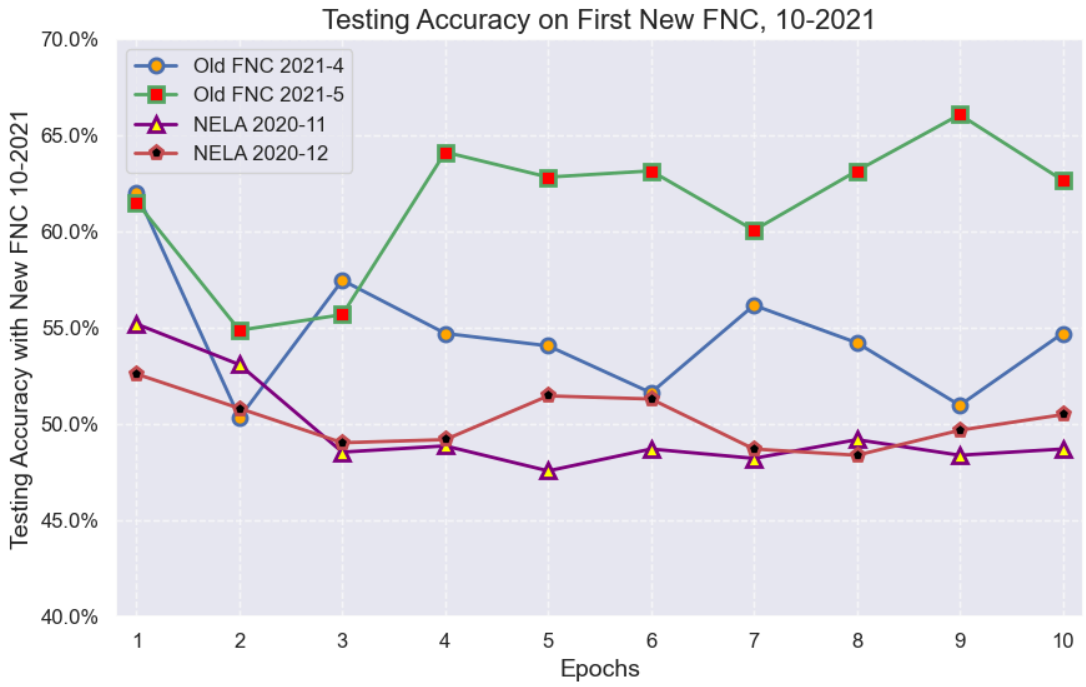


Figure 6: BERT Testing Accuracy Against the 10-2021 Dataset

This graph shows the testing accuracies of models against the 10-2021 dataset. Overall, this comparison was rather middling compared to the second dataset after as the general accuracies could only reach the mid 60s with not much more accuracy beyond. The most important note is the testing accuracy of the Old FNC datasets were much better compared to the NELA datasets.

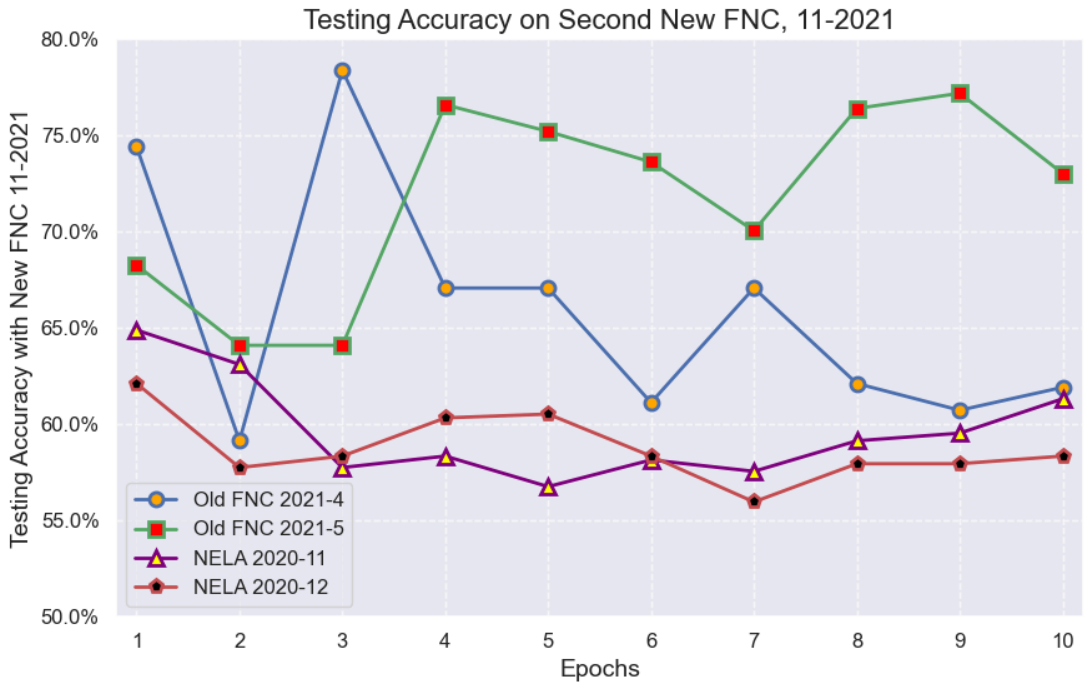


Figure 7: BERT Testing Accuracy Against the 11-2021 Dataset

This graph shows the testing accuracies of models against the 11-2021 dataset with much better accuracy compared to the previous dataset as we can observe that the models could reach rough accuracies in the 70s with the models based on the Old FNC datasets performing particularly well.

We were able to draw trends and conclude that the testing accuracies of the Old FNC datasets must be much higher due to relevancy of the labeled data. With NELA, the data is labeled more so on whether a news source can be considered reliable. While that provides a good baseline for the project, the lack of direct relevance will result in lower accuracies. The models based on the Old FNC data will be higher in testing accuracy due to their relevance to the topic of COVID-19 misinformation.

Snorkel Results

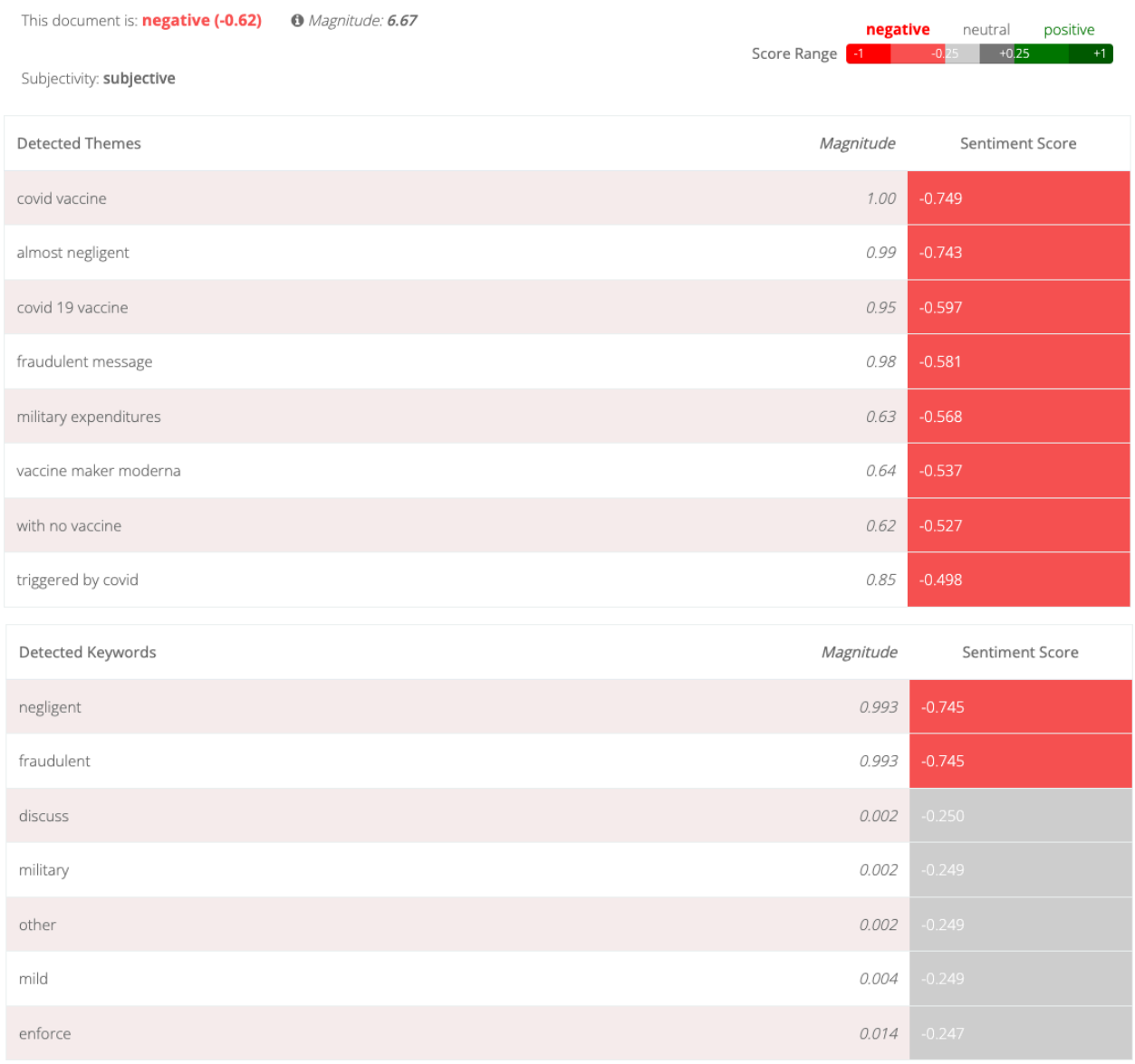
Comparison between updated and original model

There were two Snorkel models produced with the first model having decent results. However, once the Snorkel Pipeline was updated with the updated model we found that the accuracy of the updated model was 73%. This was a significant improvement on the BERT model that was used beforehand to serve as our baseline which had an average of around 55%. This was a 32.7273% increase over our baseline which is a significant improvement over the original model.

We were extremely happy with the improvement of our framework in producing datasets containing high quality labeling. With the results of such high accuracy, we believe our framework would reduce the need for hand labeling dramatically as the majority of the labeled data would be high quality already. There would be a need for physical parsing of the file to check for poor labeling, but overall the majority of the labeling will be complete. With the weak supervised produced models, there can be a mild amount of work done on the dataset which can be used to continue enhancing sophisticated machine learning models in reducing KO and focusing on identifying misinformation.

We included some visualizations of how the framework treated information. Additionally, there are visualizations of the most popular words that were selected by the model when it produced new data sets.

Visuals of Snorkel Model Output



Detected Themes	*Magnitude*	Sentiment Score
covid vaccine	*1.00*	-0.749
almost negligent	*0.99*	-0.743
covid 19 vaccine	*0.95*	-0.597
fraudulent message	*0.98*	-0.581
military expenditures	*0.63*	-0.568
vaccine maker moderna	*0.64*	-0.537
with no vaccine	*0.62*	-0.527
triggered by covid	*0.85*	-0.498

Detected Keywords	*Magnitude*	Sentiment Score
negligent	*0.993*	-0.745
fraudulent	*0.993*	-0.745
discuss	*0.002*	-0.250
military	*0.002*	-0.249
other	*0.002*	-0.249
mild	*0.004*	-0.249
enforce	*0.014*	-0.247

Figure 8: First Snorkel Sentimentality Towards Specific Words and Phrases

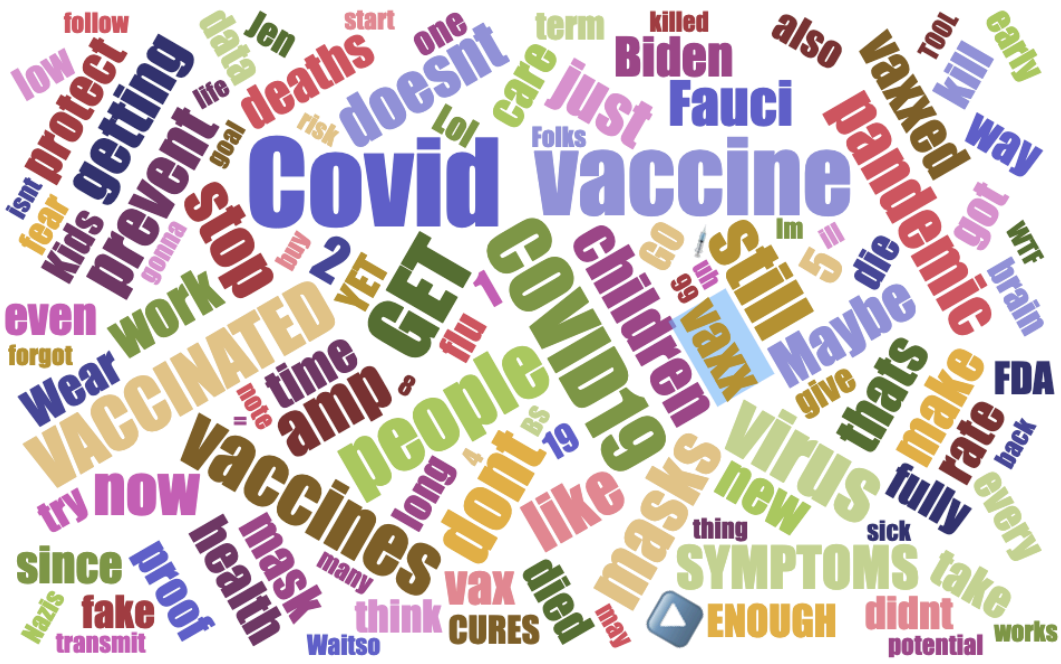


Figure 9: First Snorkel Model Word Cloud

The first word cloud covers the primary words caught such as Covid, Vaccine, Fauci, and more showing that it overall can catch target words we are searching for. However, it also shows that some words are still caught that are not fully related to misinformation such as the play emoji or “Lol”. We were happy with the general labels our framework could produce based on these models as it produces a wide variety of potential misinformation keywords related to COVID-19.

This document is: **positive (+0.82)** *Magnitude: **13.35***

Subjectivity: **subjective**

negative neutral **positive**

Score Range -1 -0.25 +0.25 +1

Detected Themes	*Magnitude*	Sentiment Score
dont delay	*0.98*	+0.731
florida concert	*0.92*	+0.700
jovi tested positive	*1.00*	+0.671
professional and personal	*0.72*	+0.589
great product	*0.99*	+0.550
doctors and hospitals	*0.68*	+0.534
covid testing locations	*1.00*	-0.750
covid clinics	*1.00*	-0.749
covid vaccine	*1.00*	-0.749
further suffering	*1.00*	-0.748
self isolate	*0.99*	-0.744
against the infection	*0.71*	-0.569
not be misled always	*0.73*	-0.555
most vulnerable	*0.64*	-0.544
covid pandemic	*0.87*	-0.543

Figure 10: Second Snorkel Sentimentality Towards Specific Words and Phrases

The takeaway from the sentimentality report is that the keywords that were identified from our output were generally more positive than that of the misinformation keywords which were generally more negative. However this still did not mean that the key words were void of negative words as shown on the graphs above, but the general contrast shows a clear distinction between true and misinformation regarding Covid 19.

Figure 11: Second Snorkel Model Word Cloud

The second word cloud produced provides a different perspective on how the self labeling model changed compared to the previous model. There is a more poor collection of data such as a higher amount of numbers and keywords such as “amp” or “UK” caught. However, the general information was still properly collected.

Future Work

While we were able to make significant progress within the 2 months of work on this project, there is still more that can be done to continue to improve our models. We discovered multiple aspects of the project that could be improved by researchers who wish to continue the progress that we made. The first improvement could be made with the provided datasets when used for training and testing. The datasets provided were typically poor with what data was labeled, the information related to the Old FNC dataset were sometimes completely unrelated to COVID-19 resulting in poorer training of our models. The NELA datasets were unrelated to COVID-19, so the poor accuracy was expected.

However, we have a hypothesis that if future research is conducted that using datasets that are all heavily related to COVID-19 would result in even better results.

Additionally, using datasets such as the Old FNC set with even more labels beyond the typical 500 entries could allow for models that are even more diverse, less probable of overfitting, and capable of achieving higher testing accuracies. This would require much greater hand labeling prior to the research to have properly labeled datasets capable of providing a robust amount of information for the machine learning models. Furthermore, the addition of more diverse models can result in potential improvements in the testing accuracy and the quality of the self labeling that was produced by our Snorkel framework. The implementation of different frameworks such as new open source models such as LLaMA could result in a higher quality labeling function due to improved accuracy.

We believe with these changes in place, then future research would be much more successful and streamlined. The enhanced datasets and more diverse set of models would allow for improved accuracies and more general weak supervision models. With these areas addressed, then we believe researchers can build upon where we left off to achieve even greater results and breakthroughs in the subject of reducing misinformation on social media related to COVID-19.

Skill Learning

Throughout the course of the project, we have developed many technical and soft skills. Overcoming various challenges such as the software bugs that occurred during the training and testing of our models and the complexities of understanding the efficacy of each model. We have compiled a list of the most important skills developed during the lifetime of the project sorted between level 1,2, and 3 skills. This project helped us grow as a team and made us learn how to go above and beyond during the project's production.

Level 1 Skills

Facts vs. Opinions: We had to work vigorously on our skills to differentiate between facts and opinions when labeling and working with COVID-19 misinformation. During the first checkpoint of our project, we had labeled 1,000+ pieces of data as either factual or misinformation. This took rigorous fact checking and analysis of every datapoint we marked. Beyond the labeling stage, we had to measure the efficacy of the various machine learning models and how well they were capable of differentiating facts against opinions. We achieved this measurement through testing strategies comparing the models against previously labeled datasets determining the accuracy as a percentage.

Programming Techniques: Our team had to work on our debugging skills with various different python libraries. There were multiple different errors that we came across and overcame. The HuggingFace machine learning library had multiple pain points particularly with their Albert model. Through multiple different debugging attempts, it was discovered that current updates with the model diminished the accuracy of the model substantially compared to different models. Furthermore, debugging different libraries associated with our training models and graphing program resulted in an improvement in how we identify and handle bugs with the various Python libraries.

Level 2 Skills

Apply Theory to Practice: Our team focused on applying our theory to practice by forming various hypotheses and producing quantitative evidence to determine whether our assumptions produced any bias. Throughout our model creation process we would create hypotheses for the potential training and testing accuracies of the models. We discovered potential errors and improvements that we could make throughout the project. For example, we discovered that one of our models would overfit through extremely high training accuracies and mediocre testing accuracies. We were then able to modify various attributes of our models to produce improved testing accuracies. Through these hypotheses we improved the quality of the models produced.

Learn and Integrate Multiple Platforms: With the various models, we were able to implement them with Snorkel to combine their greatest strengths in a

weak supervision model. With our best models working in tandem they reduce the weak points of a singular model in the self learning model. With multiple different models in use, we garnered skills in combining the models in an effective manner for Snorkel.

Implement Functionality by Augmentation: We optimized the efficiency of our machine learning models through augmentation of their configurations. With multiple changes to various aspects such as the learning rate, number of epochs, batch size, and many other factors to maximize the efficiency of the different models. Through these changes we expanded our skills on taking premade assets and augmenting them to optimize their efficiency for text based datasets such as the COVID-19 tweets.

Level 3 Skills

Trend Recognition: Our team developed skills in visualizing the data produced by the various models created through the project's lifetime. After multiple iterations of creating the best visualization software to represent the data as we believed to be best fit, we created graphs that could represent the possible trends of every model and configuration and determine what changes could be made to improve their trends. With the addition of the hypothesis, the visualization, and their analysis resulted in remarkable improvements to future results of the different models. In addition we were able to see the trends of this project in terms of further improvements. This project allowed us to identify the current trends of our work and take this knowledge to help prepare for future work that can continue to identify misinformation more accurately and efficiently in the future.

Distribution Recognition: We would determine the distributions of our models through graphical visualizations and modify our predicted trends for future modifications. The distributions we took into consideration were the training accuracy, testing accuracy, and their differences between models including their specific configurations. We delved even further by consistently producing and predicting trends of our future models as well.

References

Aimpoint Digital Partners With Snorkel AI to Accelerate AI Development: The partnership between noted analytics firm Aimpoint Digital and data-centric AI innovator Snorkel AI aims to accelerate the adoption of machine learning for Fortune 500 enterprises. (2022). In *NASDAQ OMX's News Release Distribution Channel*. NASDAQ OMX Corporate Solutions, Inc.

Bach, S. H., Rodriguez, D., Liu, Y., Luo, C., Shao, H., Xia, C., Sen, S., Ratner, A., Hancock, B., Alborzi, H., Kuchhal, R., Ré, C., & Malkin, R. (2019). Snorkel DryBell: A Case Study in Deploying Weak Supervision at Industrial Scale. *Proceedings of the 2019 International Conference on Management of Data*, 362–375. https://doi.org/10.1145/3299869.3314036

Barthel, M. (2016, December 15). *Many Americans believe fake news is sowing confusion*. Pew Research Center. https://www.pewresearch.org/journalism/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/

Bolla, B. K., Pattnaik, S. R., & Patra, S. (2024). Detection of Objectionable Song Lyrics Using Weakly Supervised Learning and Natural Language Processing Techniques. *Procedia Computer Science*, *235*, 1929–1942. https://doi.org/10.1016/j.procs.2024.04.183

Davies, J. (n.d.). *Word Cloud Generator*. Word cloud generator. https://www.jasondavies.com/wordcloud/

Farchi, A., Laloyaux, P., Bonavita, M., & Bocquet, M. (2021). Using Machine Learning to Correct Model Error in Data Assimilation and Forecast Applications. https://doi.org/10.5194/egusphere-egu21-4007

Gupta, D. (2020). Alumni launch Snorkel, rethink machine learning. In *University Wire*. Uloop, Inc.

Mallory, E. K., de Rochemonteix, M., Ratner, A., Acharya, A., Re, C., Bright, R. A., & Altman, R. B. (2020). Extracting chemical reactions from text using Snorkel. *BMC Bioinformatics*, *21*(1), 217–217. https://doi.org/10.1186/s12859-020-03542-1

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid Training Data Creation with Weak Supervision. *Proceedings of the*

VLDB Endowment, *11*(3), 269–282. https://doi.org/10.14778/3157794.3157797

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2020). Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, *29*(2–3), 709–730. https://doi.org/10.1007/s00778-019-00552-1

Ratner, A., Hancock, B., Dunnmon, J., Goldman, R., & Ré, C. (2018). Snorkel MeTaL: Weak Supervision for Multi-Task Learning. *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, *2018*, 1–4. https://doi.org/10.1145/3209889.3209898

Ré, C. (2018). Software 2.0 and Snorkel: Beyond Hand-Labeled Data. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2876–2876. https://doi.org/10.1145/3219819.3219937

Sentiment analysis report. Sentiment Analysis - Free Online Demo. (n.d.). https://text2data.com/Demo?lnk=214207258

Snorkel AI Teams with Microsoft. (2023). In *Wireless News*. Close-Up Media, Inc.