**IT 6903 Project Phase 2 – Proposal:  
COVID-19 Social Media Posts Multiclass Classification with Deep Learning**

Elaine Harris

Kennesaw State University

MSIT Program – Fall 2020

Contact: [eharr147@students.kennesaw.edu](mailto:eharr147@students.kennesaw.edu)

Table of Contents

[Abstract 3](#_Toc51327131)

[**IT 6903 Assignment 1 Topic Modeling With Latent Dirichlet Allocation (LDA)** 4](#_Toc51327132)

[Real Life Application 4](#_Toc51327133)

[Social Media Posts Format 5](#_Toc51327134)

[Proposed Solution 6](#_Toc51327135)

[Topic Modeling with Latent Dirichlet Allocation 7](#_Toc51327136)

[LDA Practical Example 8](#_Toc51327137)

[Planned Execution and Expected Results 9](#_Toc51327138)

[Additional Readings 10](#_Toc51327139)

[References 11](#_Toc51327140)

[Figure 1 LDA Model Code Snippet 9](#_Toc51327177)

[Figure 2 Coherence Model Code Snippet 9](#_Toc51327178)

Abstract

This project proposal follows up on the application proposed in Assignment 1: Topic Modeling with Latent Dirichlet Allocation. The domain is the collection of Facebook and Twitter posts by the Singapore Ministry of Health related to the COVID-19 Pandemic between February and July 2020.

The proposed LDA analysis was in fact executed, but the resulting topics do not provide a clear distinction between the identified topics. Furthermore, it failed to single out particular subjects such as social distancing, facial coverings, etc. For this project, we will manually label a portion of the posts in the dataset and attempt to generate labels for the rest of the posts using Multiclass Classification Algorithms in a supervised learning approach. We will implement the algorithms in Python, using Keras and Pytorch libraries.

**IT 6903 Project Phase 2 – Proposal: COVID-19 Social Media Posts Multiclass Classification with Deep Learning**

This project proposal attempts to improve on the results of the LDA Topic Modeling Analysis suggested by the author in the Assignment 1 paper submitted for this course: *IT 6903 Assignment 1: Topic Modeling With Latent Dirichlet Allocation (LDA).*

In the Assignment 1 paper, the author proposes the use of LDA algorithm to discover topics in the Social Media Posts by the Singapore Ministry of Health. This work is related to the 7993 Capstone class project *Analyzing Risk Communication and Behavioral Change During COVID-19 Pandemic*.

Since the submission of the Assignment 1 paper, the proposed implementation has indeed been executed. The final model was implemented using the Gensim library wrapper of the Mallet LDA Algorithm. A total of 15 topics were created by the algorithm, and the Dominant Topic for each post was identified by selecting the topic with the highest percentage contribution.

Upon visual examination, it was determined that while some topics have a clearly identified theme such as New Cases, Patient Condition Updates, etc, many others have a few posts with a high matching rate to the Dominant Topic, and many other posts with low matching rate to each topic overall. This distribution makes the sets look messy and confusing, and makes it hard to distinguish topics we are interested in, such as Social Distancing, Masks, Hygiene, etc.

In this **IT 6903 Project**, we propose to improve the analysis of the dataset posts by applying a supervised Deep Learning approach to generate *Multiclass Labels*. The labels will be manually created for a subset of the posts to use as the training data. We will then run two or three different Deep Learning models in Keras and compare the results produced on the unlabeled posts.

Finally, if the results of the Multiclass Classification algorithm are superior to the topics obtained in the LDA Algorithms, the new labels will be used in the 7993 Capstone project. However, the analytics and results of the Capstone project are out of scope for this proposal.

# 

# Dataset and LDA Algorithm Results

The following is an excerpt from the Capstone Project Requirements by Dr Shirley Tian:

***Project Title: Analyzing Risk Communication and Behavioral Change During COVID-19 Pandemic***

*Introduction: We are particularly interested in how to track, model, understand and predict the spread of COVID-19 through data mining and visualization methods. We will address the following questions: What social behaviors should policymakers and government leadership take into consideration when engaging in public health messaging? How can messaging better promote facts and counter misinformation? We will compare both national and international public health messaging and their effects by analyzing and visualizing social media data.*

## Social Media Posts Format

This project only collects posts from the Singapore Ministry of Health, and the text is very repetitive. We need to analyze the text and attempt to identify topics related to social distancing, facial coverings and hygiene, Government-imposed restrictions, at a minimum, and any other topics that may surface after analysis.

Below are some examples of real posts collected from Twitter and Facebook.

|  |  |  |
| --- | --- | --- |
| **Username** | **Text** | **Date** |
| sporeMOH | There are currently 117 confirmed cases who are still in hospital. Of these, most are stable or improving, and none is in the intensive care unit. | 2020-08-01 |
| sporeMOH | As of 1 August 2020, 12pm, 249 more cases of COVID-19 infection have been discharged from hospitals or community isolation facilities. In all, 46,740 have fully recovered from the infection and have been discharged. | 2020-08-01 |
| sporeMOH | Of the new cases, 99% are linked to known clusters, while the rest are pending contact tracing. | 2020-08-01 |
| sporeMOH | As of 1 August 2020, 12pm, we have confirmed and verified an additional 307 cases of COVID-19 infection in Singapore. Breakdown: 5 imported, 1 case in the community & 301 cases residing in dorms. | 2020-08-01 |
| sporeMOH | As of 1 August 2020, 12pm, we have preliminarily confirmed an additional 307 cases of COVID-19 infection in Singapore. | 2020-08-01 |
| sporeMOH | There are currently 136 confirmed cases who are still in hospital. Of these, most are stable or improving, and none is in the intensive care unit. | 2020-07-31 |

Table Sample Twitter Posts

|  |  |
| --- | --- |
| Date | Post |
| 5/1/2020 | As of 1 May 2020, 12pm, we have confirmed and verified an additional 932 cases of COVID-19 infection in Singapore. The breakdown is as follows:  a) Imported cases: 0 b) Cases in the community: 11 (5 Singaporeans/Permanent Residents, 6 Work Passes) c) Work Permit holders (residing outside dormitories): 16 d) Work Permit holders (residing in dormitories): 905 Of the new cases, 70% are linked to known clusters, while the rest are pending contact tracing.  24 more cases of COVID-19 infection have been discharged from hospitals or community isolation facilities. In all, 1,268 have fully recovered from the infection and have been discharged from hospitals or community isolation facilities.  There are currently 1,764 confirmed cases who are still in hospital. Of these, most are stable or improving, and 23 are in critical condition in the intensive care unit. Read more in the press release: https://www.moh.gov.sg/news-highlights/details/24-more-cases-discharged-932-new-cases-of-covid-19-infection-confirmed |
| 5/1/2020 | As of 1 May 2020, 12pm, we have preliminarily confirmed an additional 932 cases of COVID-19 infection in Singapore, the vast majority of whom are Work Permit holders residing in foreign worker dormitories. Five cases are Singaporeans/ Permanent Residents. We are still working through the details of the cases, and further updates will be shared via the MOH press release that will be issued tonight. https://www.moh.gov.sg/news-highlights/details/932-new-cases-of-covid-19-infection |
| 5/2/2020 | <One More Week of Tightened Circuit Breaker Measures; Gradual Easing Thereafter> We tightened the circuit breaker measures two weeks ago and said we would review the situation by 4 May. To date, we have seen a reduction in our daily community infection numbers. But we are still not in single digits. So aside from some minor adjustments (for TCM practitioners and what one can do within the grounds of strata-titled residential developments), we will have to continue with the tight measures for at least another week. From 12 May, we can allow some gradual easing of these tightened measures. In particular, we will allow selected activities and services to resume operations, including home-based businesses, selected food manufacturing and retail outlets; laundry and barber services, and pet supplies stores.  We will also phase in the reopening of the economy, and allow more people to resume going to work. The key precondition for this to happen is stricter requirements in the workplace for safe distancing, with employers to take ownership and responsibility of these measures. The specific measures are being worked out, and government agencies will be engaging industry associations, business chambers and firms in the coming days.  One specific requirement is that all businesses must put in place the SafeEntry app to log the entry and exit of their staff and visitors. This information will enable us to speed up contact tracing.  Further down the road, from 19 May, we will start to bring back students in small groups for face-to-face lessons. We will focus on the graduating cohort who are taking national exams, and amongst them, priority will be given to those who require school facilities for coursework and practical sessions, and those who need additional support during the school vacation period. The Institutions of Higher Learning especially the ITEs will also bring back small groups of students on campus for critical consultations, projects or practicums We have provided a broad outline of how we intend to adjust the measures in the coming weeks so that everyone knows what to expect. But the situation is fluid, and the measures may have to be further adjusted along the way.  Most importantly, remember that the circuit breaker is still in place till 1 June. This is not the time to slacken and let our guard down. We must continue to stay disciplined and vigilant. Stay home as much as possible. Go out for essential activities only, and when you have to do so, try to go out one person at a time. We still have a long fight ahead of us, and the virus can flare up again anytime. Let’s stay focused, work together and win this fight against the virus. |

Table Sample Facebook Posts

## Latent Dirichlet Allocation (LDA) Results

Original posts were broken down into sentences, cleaned up and lemmatized. We selected to lemmatize trigrams and use only nouns.

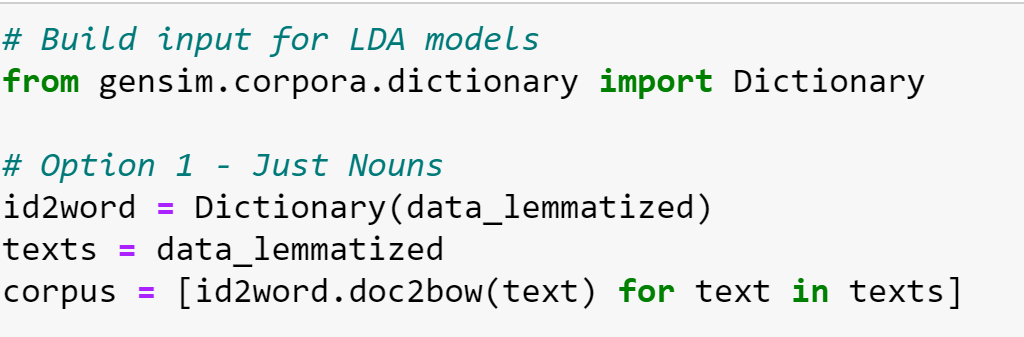


Figure 1 Input for LDA Model

Next, several models were created in order to select the optimum number of topics for the final model. We selected to use 15 topics.

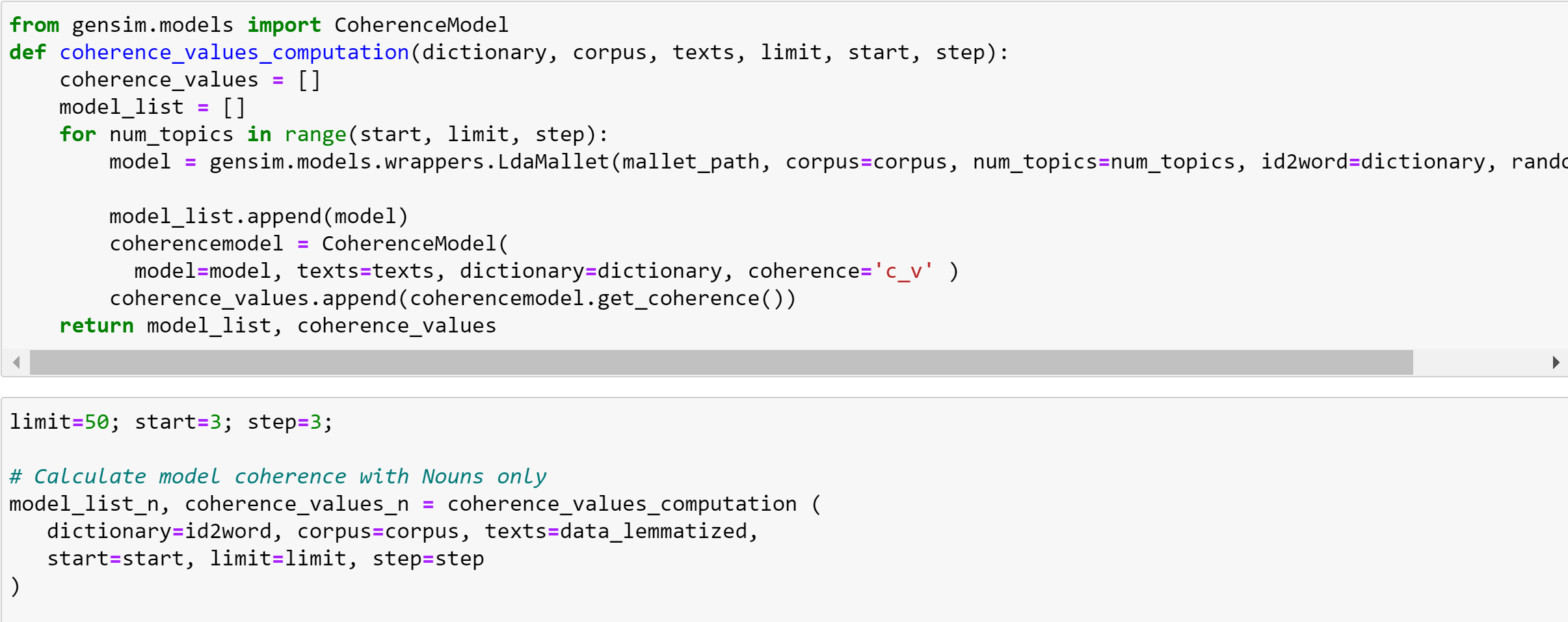


Figure 2 Calculate Coherence Score for up to 50 Topics

# 

Figure 3 Coherence Score Graph

The topics identified by the algorithm are listed in *Figure 4*. Only the top 10 keywords and their probability are represented in this table. The topic Description was inferred based on the content of the posts that more closely match the topic based on the highest probability value for the dominant topic.



Figure 4 LDA Topics - Keywords and Probability Distribution

## LDA Topics Lack Clarity

After topics were identified and given a collective description, visual observation easily determines that the posts that more closely match the dominant topic appear to be representative of the subject. However, many posts match all topics at low probabilities, and even the highest match (the dominant topic) does not appear to fit with the other posts. *Figure 5* shows an example for *Topic 9 – Medical Subsidies*. The highlighted posts match the topic description, but there are multiple other posts that are not a good match. This same issue appears on most of the topics identified by the LDA Analysis.

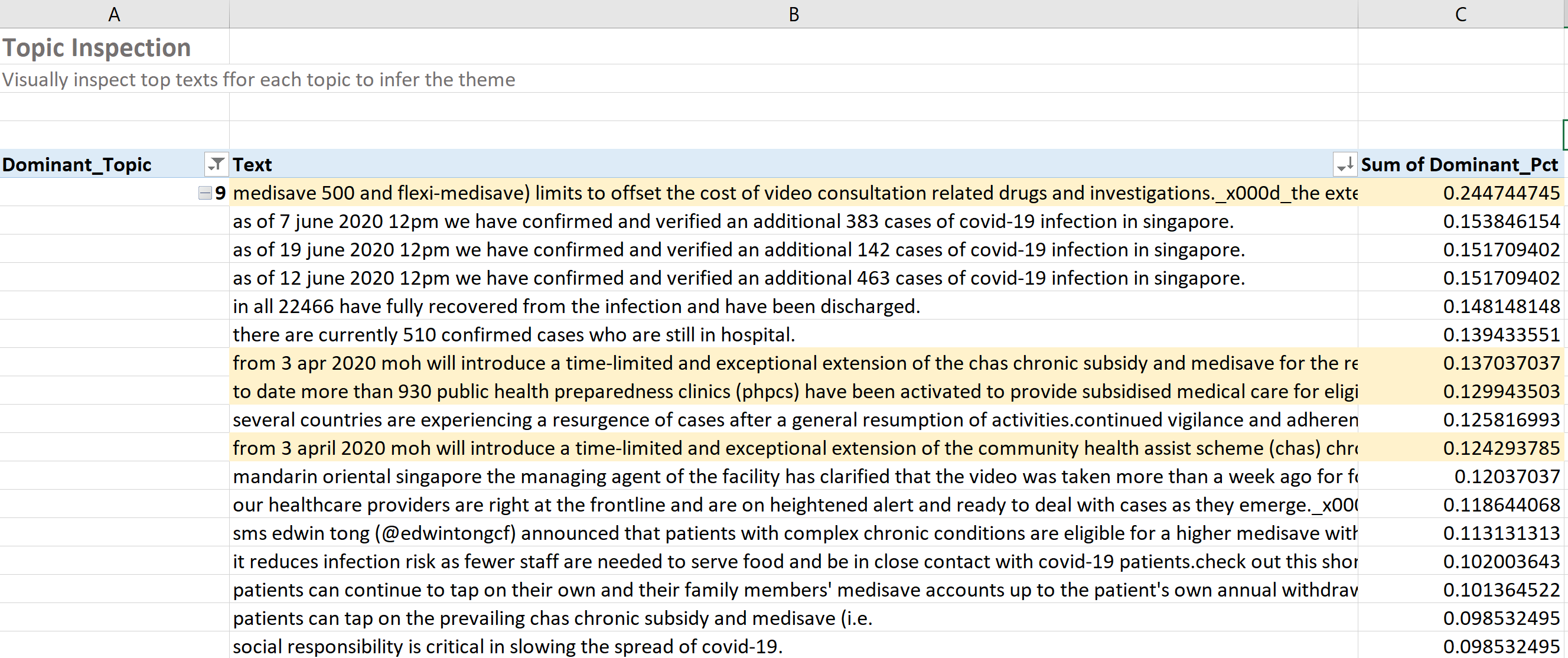


Figure 5 Posts under Topic 9 - Medical Subsidies

# Proposed Solution

We will use a different approach to identify the topics in the text. We will label part of the dataset (to be used as the training data) and use two or three variations of Deep Learning Multiclass Classification algorithms to label the rest of the data.

The training data will be labelled manually, and will be loosely based on the topics identified in LDA. The output of the LDA analysis is already broken down into sentences. We will run the dataset through text cleanup code that emulates the text preparation used to train the GloVe embeddings, which will be used as the word embedding layer in our Deep Learning Models.

Deep Learning algorithms will be implemented based on guidance found in the following article: *Using Deep Learning for End to End Multiclass Text Classification* (Agarwal, 2020). Another two articles that contain relevant code and recommendations might also be used: Multi-*Class Text Classification with LSTM* (Li, 2020) and *Python for NLP: Multi-label Text Classification with Keras* (Malik, 2020). The last article is on Multi-label classification, meaning one post might match one or more categories. At this point, we prefer to use Multiclass classification and obtain a single label per post, but we will keep in mind that Multi-label classification might also be an option to solve our problem.

## Deep Learning Model Architecture

Based on the guidelines in the articles mentioned earlier, the model will have the following basic layers:

1. Word Embedding Layer: we will use pre-trained GloVe Embeddings[[1]](#footnote-1). This layer encodes words as number vectors that can be processed by the algorithm. There are many options to create the embedding; we will use GloVe because it is a popular pre-trained model among NLP practitioners that is fairly easy to use.
2. LSTM Layer: executes the text classification. “Long Short Term Memory networks (LSTM) are a subclass of RNN[[2]](#footnote-2), specialized in remembering information for extended periods…which is pretty useful in text classification tasks” (Agarwal, 2020)
3. Output Layer: must create one output for each labelled class.
4. Activation Function: softmax[[3]](#footnote-3) for multi-class classification (Li, 2020). Agarwal uses RELU[[4]](#footnote-4) for his BiDirectional RNN. (Agarwal, 2020)
5. Finally, Li uses categorical cross-entropy as the loss function (Li, 2020) Categorical crossentropy is a “loss function that is used in multi-class classification tasks. These are tasks where an example can only belong to one out of many possible categories, and the model must decide which one.” ("Categorical crossentropy loss function", 2020)

# RESULTS FOR CAPSTONE REPORT

Execution is detailed in Excel file Step7\_Consolidate\_Classified\_Text.xlsx, Read Me tab.

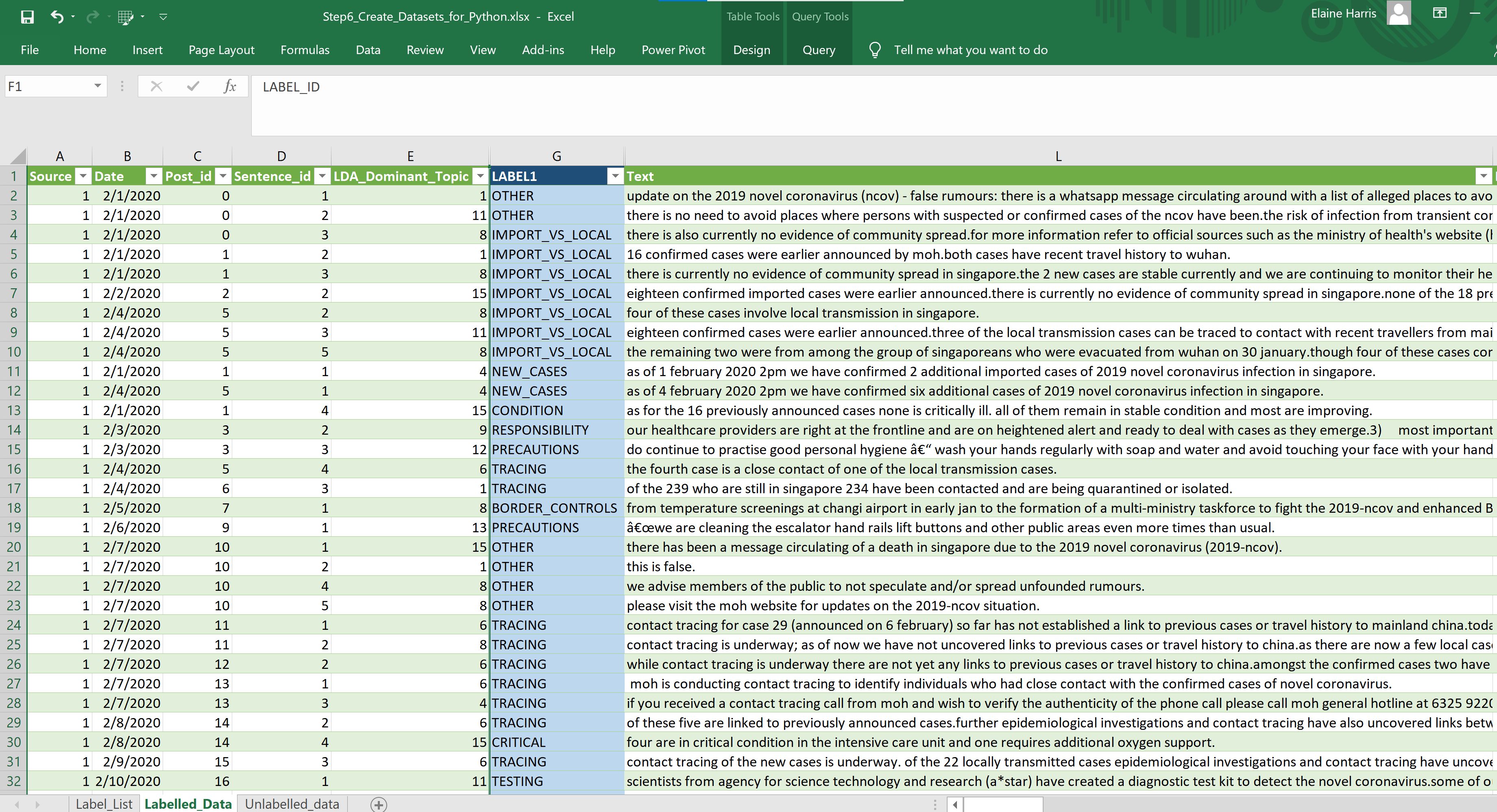
Steps:

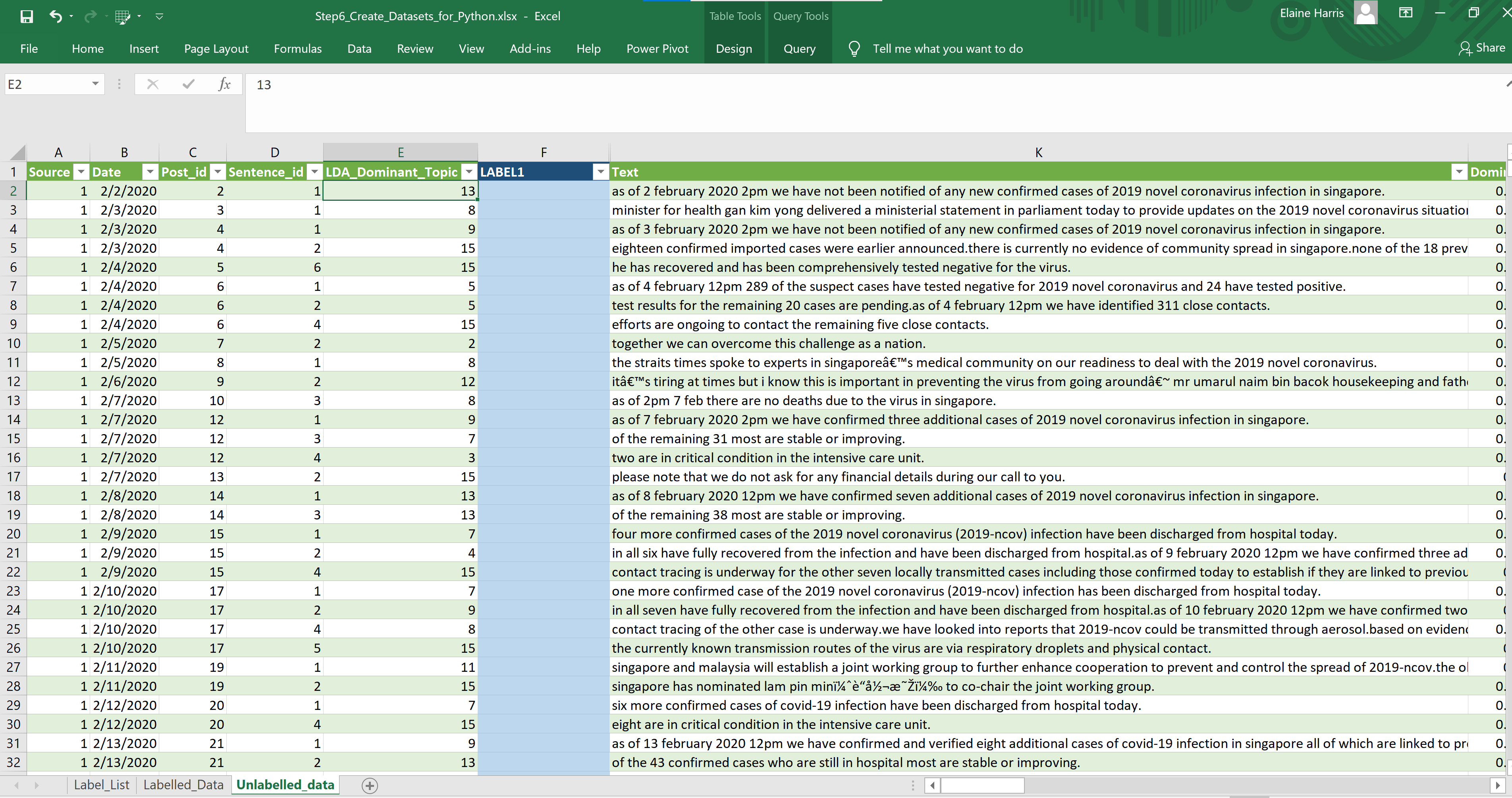
1. Identify categories based on the topics identified by LDA algorithm.

Topics are ranked from most to least interesting: Precautions is the most interesting topic, while statistics on number of recovered patients and percentage of cases traced to clusters are the least interesting. 

(original table in Step7\_Consolidate\_Classified\_Text.xlsx)

1. Label several posts, leaving other similar posts unlabelled. Split the dataset into Labelled and Unlabelled. The labelled dataset will be used to train the model, and the unlabelled dataset will be fed to the model to classify remaining posts. Split is roughly 30% training, 70% unlabelled. Datasets are saved to file Step6\_Create\_Datasets\_for\_Python.xlsx. Examples of Labelled and Unlabelled datasets below:





1. Create the Deep Learning Multiclass Classification Model in Python. Jupyter Notebook located in teams under Data Analytics\Topic modeling Python folder. (Shoould move this to GitHub).

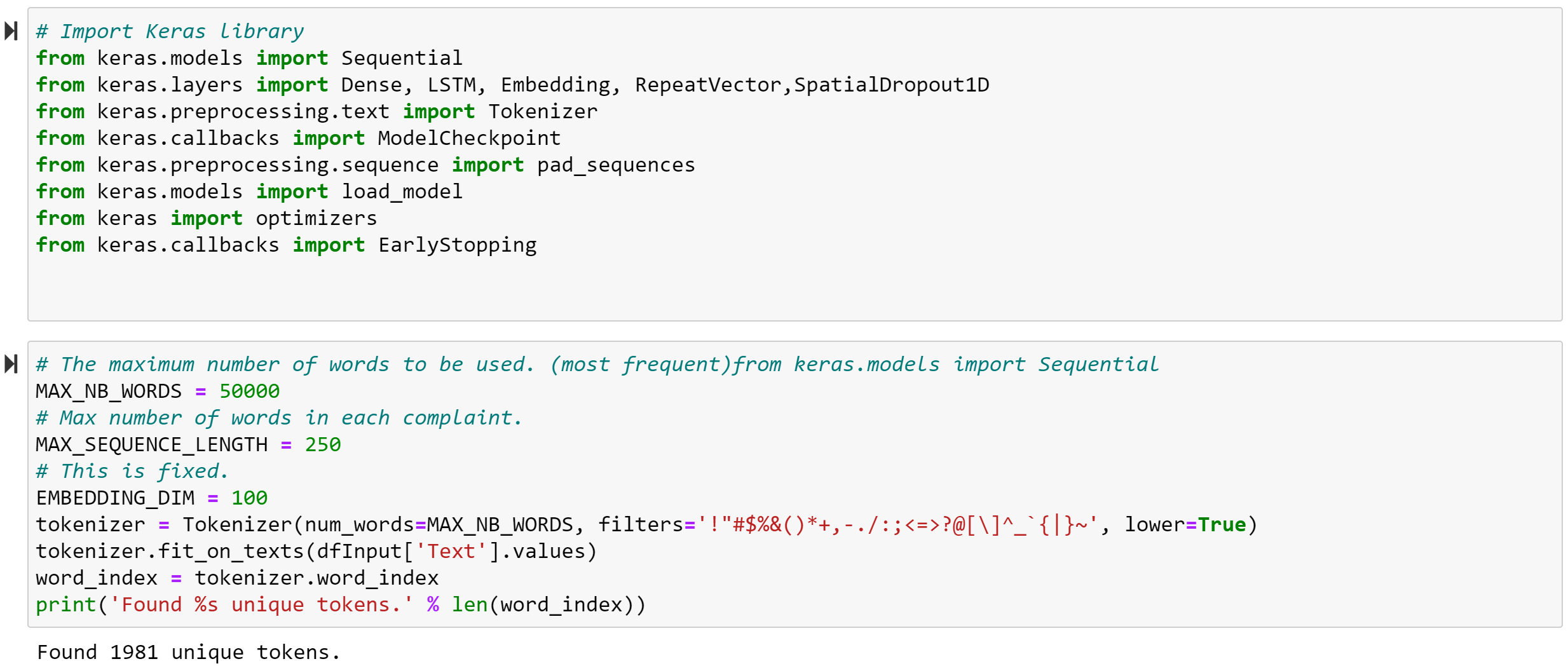
Script: multiclass-model-01.ipynb

Classified dataset: Combined\_sentences\_multiclass\_lstm.csv

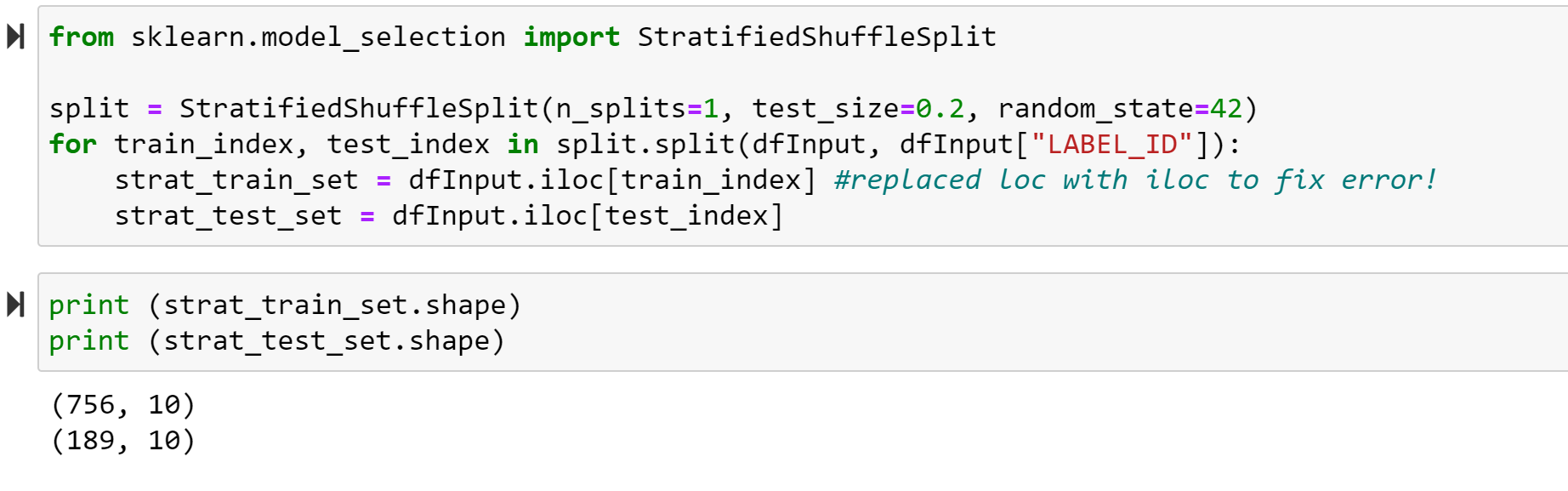
Execution in Python:

(Add to references:

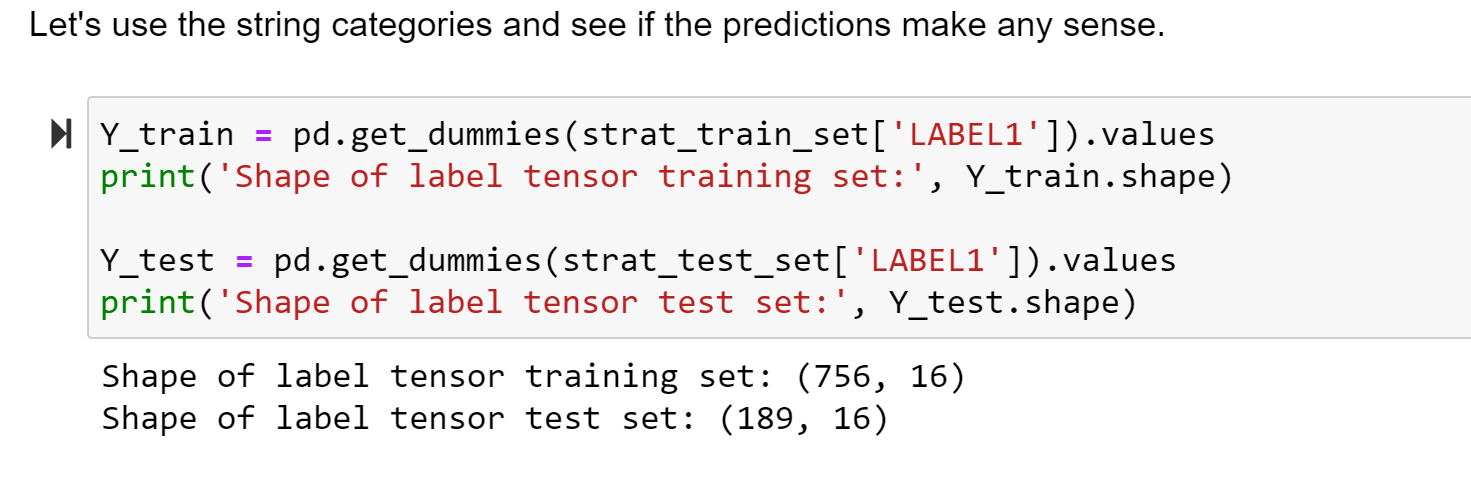
Code in this Notebook is based on this tutorial: <https://towardsdatascience.com/multi-class-text-classification-with-lstm-1590bee1bd17>)



Use SkLearn Startified Shuffle to ensure all labelled categories will be present in the Training Set:



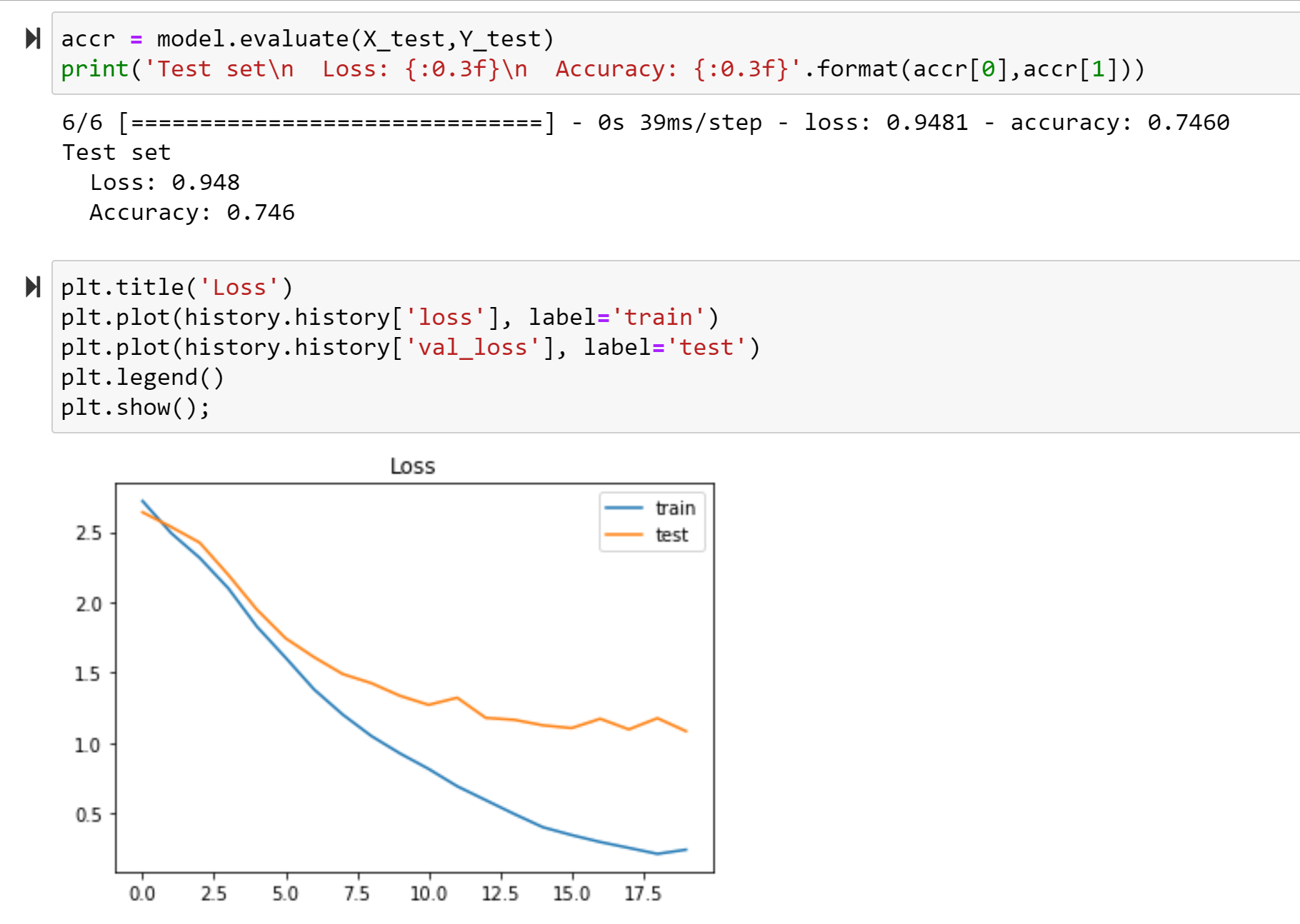
Create One Hot Encode for the labels and execute the model:



Model training:



Final Model Performance:



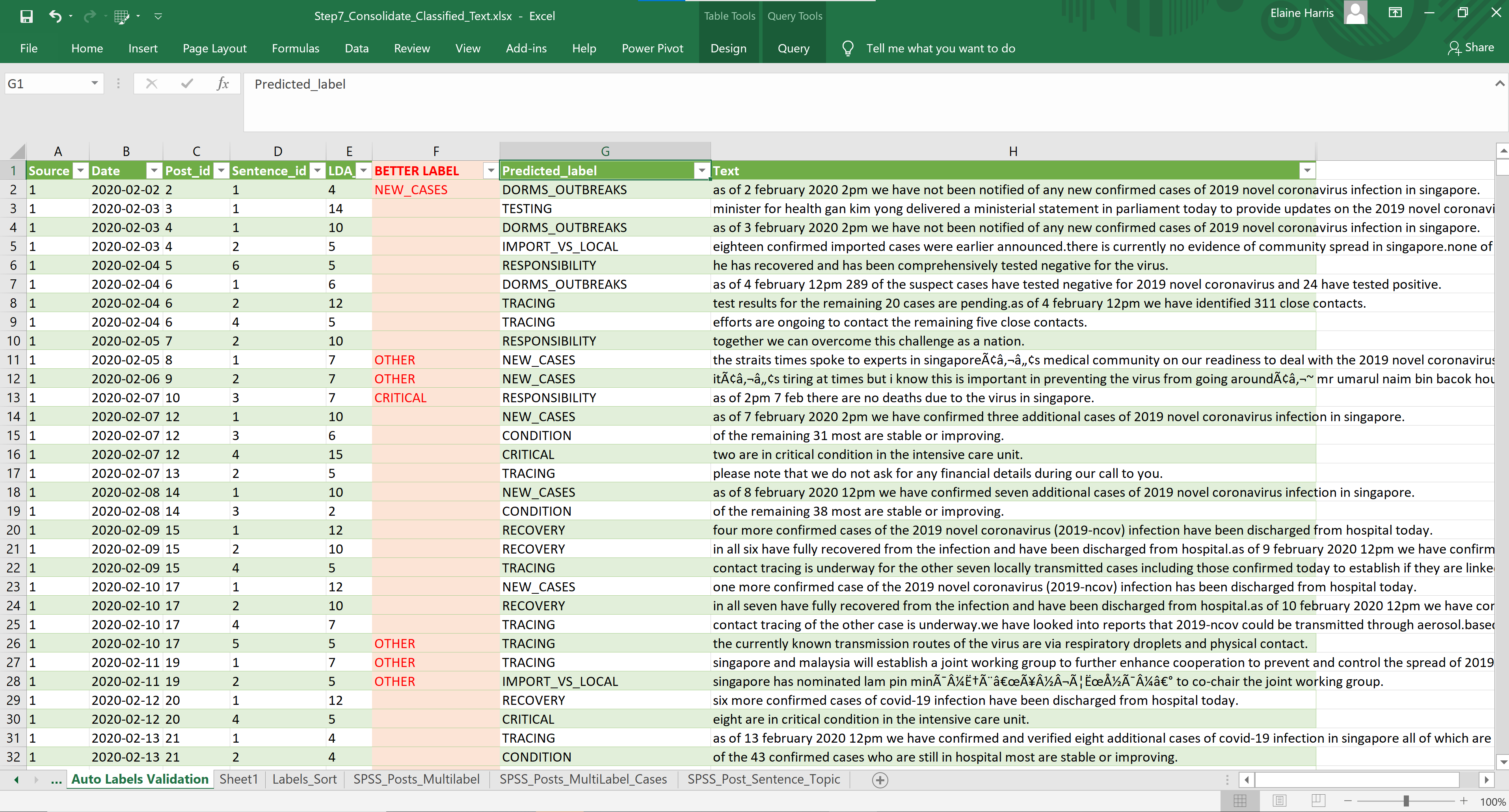
1. Visually inspect the classified posts, and make any adjustments deemed necessary by manually identifying a better match. Power Query in Excel is then used to perform the label substitution and consolidate the previously labelled dataset and the output of the Classifier into a single dataset for further analytics.

(data below comes from Step7\_Consolidate\_Classified\_Text.xlsx)

Performance of the Classifier:

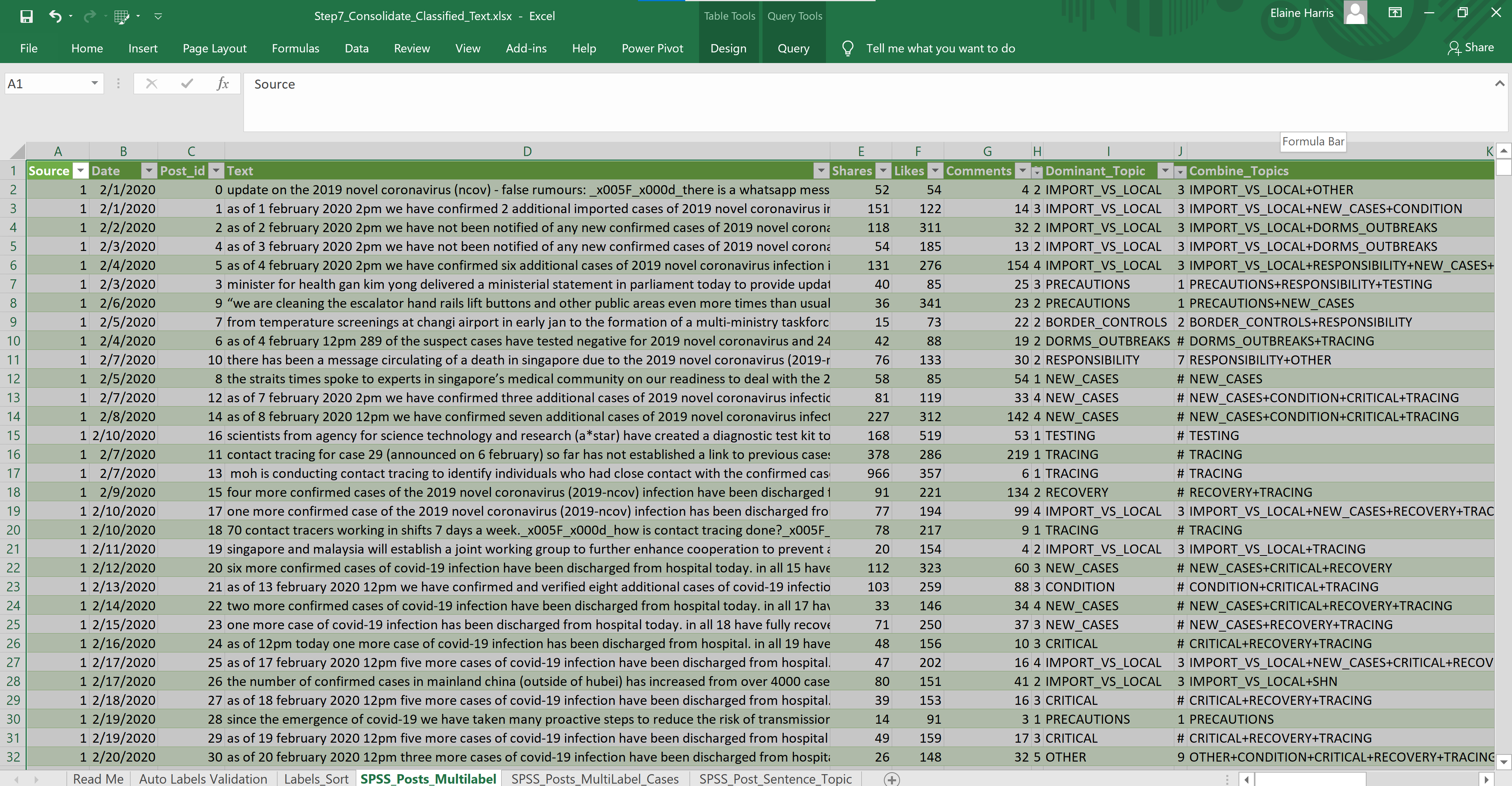
|  |
| --- |
| Deep Learning LSTM Model with Softmax Activation Function for Multiclass Classification. Loss function is CrossEntropy. |
| Model was trained in 20 epochs. Performance to Testing data: loss: 0.9481 - accuracy: 0.7460 |

Examples of manual label substitution:



1. Create a Dominant Topic category for each post. A single post may have multiple topics. The lowest ranked (most interesting) post is provided as the Dominant Topic for the Post. Other applicable topics are also provided, ranked by importance, however, most if the data analysis will rely on the Dominant Topic.

Dataset is published in file Step7\_Consolidate\_Classified\_Text.xlsx, sheet *SPSS\_Posts\_Multilabel*.



References

Agarwal, R. (2020). Using Deep Learning for End to End Multiclass Text Classification. Retrieved 7 October 2020, from <https://towardsdatascience.com/using-deep-learning-for-end-to-end-multiclass-text-classification-39b46aecac81>

Brownlee, J. (2020). A Gentle Introduction to the Rectified Linear Unit (ReLU). Retrieved 7 October 2020, from <https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/>

Brownlee, J. (2020). A Gentle Introduction to the Rectified Linear Unit (ReLU). Retrieved 7 October 2020, from <https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/>

GeeksforGeeks. (2020). Introduction to Recurrent Neural Network - GeeksforGeeks. Retrieved 7 October 2020, from <https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/>

Li, S. (2020). Multi-Class Text Classification with LSTM. Retrieved 7 October 2020, from <https://towardsdatascience.com/multi-class-text-classification-with-lstm-1590bee1bd17>

Malik, U. (2020). Python for NLP: Multi-label Text Classification with Keras. Retrieved 7 October 2020, from <https://stackabuse.com/python-for-nlp-multi-label-text-classification-with-keras/>

Pennington, J. (2020). GloVe: Global Vectors for Word Representation. Retrieved 7 October 2020, from <https://nlp.stanford.edu/projects/glove/>

Softmax function. (2020). Retrieved 7 October 2020, from <https://en.wikipedia.org/wiki/Softmax_function>

**END OF DOCUMENT**

1. GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. (Pennington, 2020). [↑](#footnote-ref-1)
2. Recurrent Neural Network(RNN) are a type of Neural Network where the output from previous step are fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words. (GeeksforGeeks, 2020) [↑](#footnote-ref-2)
3. Softmax Function is “is a generalization of the logistic function to multiple dimensions. It is used in multinomial logistic regression and is often used as the last activation function of a neural network to normalize the output of a network to a probability distribution over predicted output classes.” ("Softmax function", 2020) [↑](#footnote-ref-3)
4. The rectified linear activation function or ReLU for short is a piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero. It has become the default activation function for many types of neural networks because a model that uses it is easier to train and often achieves better performance. (Brownlee, 2020) [↑](#footnote-ref-4)