[Colab link]

Eric Liang, Avi Walia
All work done jointly

Part I

A. After cleaning, the data contains 18,042 positive sentiments, 15,397 negative sentiments, and 7,712 neutral sentiments.
D. We may want to keep punctuation since punctuation sometimes changes the meaning of phrases. For example, “Let’s eat, grandma” and “Let’s eat grandma” have wildly different meanings.
E. We use the Porter stemmer.
G. The length of vocabulary is 74,221.
H. The training accuracy is 82%, and the test accuracy is 67%. The top words and counts are

Negative sentiment	Neutral sentiment	Positive sentiment
coronaviru: 6737.0 covid19: 4610.0 price: 4347.0 food: 3638.0 thi: 3225.0	coronaviru: 3812.0 covid19: 2566.0 store: 1588.0 supermarket: 1441.0 price: 1365.0	coronaviru: 7511.0 covid19: 5681.0 store: 3917.0 thi: 3781.0 price: 3339.0

I. Since ROC curves are specific for binary classification problems, it would not be applicable here.
J. The training accuracy is 73%, and the test accuracy is 62%.
K. The training accuracy is 73%, and the test accuracy is 61%. Surprisingly, with lemmatization, the accuracy is somewhat lower.

Bonus: The Naive Bayes model is a generative model since it considers a prior.

Part II

(see next page)

Finding Trending Topics about Toronto using Latent Dirichlet Allocation

I. Problem Description and Motivation

Our objective is to find the most trending topics related to Toronto by analyzing recent tweets that mention "Toronto" or use the hashtag "#Toronto". This question is significant because it can provide insights into the current interests, concerns, or events that capture people's attention in or related to Toronto. Understanding trending topics can offer valuable information for various stakeholders, including businesses looking for marketing insights, policymakers interested in public opinion, and researchers studying social dynamics. This question is challenging due to the vast amount of data on Twitter and the dynamic nature of trends. We are trying to capture very recent activity that changes day to day, sometimes even every hour. Previous studies have applied topic modelling to social media data for trend analysis, but our focus on Toronto provides a novel dataset and potentially unique insights. By employing LDA, we aim to contribute to the understanding of regional trend analysis on social media platforms.

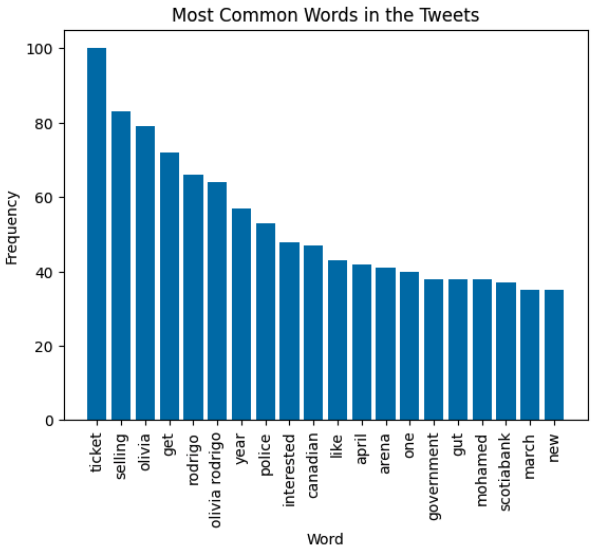
II. Describing the Data

We used the Tweepy Python library with Professor Nobre's credentials to extract tweets from the past seven days using ‘Toronto OR #Toronto’ as our search query. This query was chosen to capture the most current and relevant discussions about Toronto. We did not limit our query by anything else (such as excluding retweets or further narrowing our search by adding new keywords) because we wanted to find the most trending topics currently. Our dataset consists of 1000 tweets, which were collected on March 29th at 12:45 am. These are the 1000 most recent tweets related to Toronto. We saved this to a csv file which can be accessed from Github. For this analysis, we are only interested in the tweet text, so we dropped the other columns. Previous works have applied LDA to social media datasets for trend analysis, showing its effectiveness in

identifying dominant themes within large samples. However, our focus on a specific geographic location, in this case, Toronto, distinguishes our dataset and potential findings. The dynamic nature of trends does not allow us to compare results to other research works since time and topic are very important factors. There might be some constants (I expect things like Toronto sports teams or politics to be trending no matter the time), but most trending topics would be different. Some limitations of our data are the seven-day extraction window and the rate limit of the free version of the Twitter API. Our sample size of tweets was quite small, and while we did not expect to be able to conduct an analysis of the full population, we fear that our sample is not large enough to be able to accurately predict trends about the whole population. The short timeframe in which the data was collected severely limits our ability to discover long-term trends. The dataset's strength lies in its timeliness and relevance to current events in Toronto.

III. Exploratory Data Analysis (EDA)

We began by cleaning the data, including removing URLs, mentions, and non-alphabetic characters from the tweets. We also removed punctuation since things like tone do not matter since we care more about the actual content of the tweets, not what people are saying about the content. We then tokenized the text, lemmatized the tokens, and removed stop words. Finally, we generate bigrams on our tokens. Bigrams are two tokens frequently occurring together in the document. An example of one from our data would be “Olivia Rodrigo” since many people were tweeting about her recent concert. While our model would treat these words as separate, and most likely group them into



the same topic, it still helps us when looking at topics and seeing these bigrams to be able to discern what each topic is about. We created this visualization that looked at the most common words that occur across the tweets. We can see that the words that occur the most include ticket, selling, Olivia Rodrigo, police, and others. We then conducted an LDA analysis to try to group these words with related words to determine the overall topics that are in the dataset.

We then vectorized the processed tweets into a bag-of-words (BoW), which represents tweets as vectors of word counts, ignoring the order of words but maintaining their frequency.

Finally, we split our data into training, validation, and test sets with a 70-15-15 split. We use the training set to train the model, the validation set for hyperparameter tuning, and the testing set for testing.

IV. Machine Learning Model

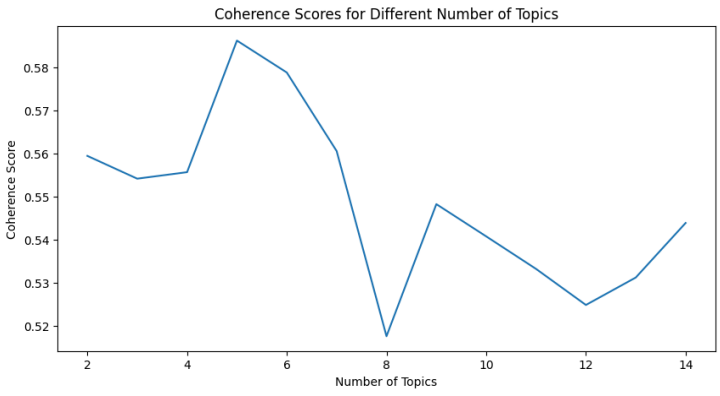
Latent Dirichlet Allocation (LDA) is a type of unsupervised learning used for topic modelling. It assumes each document (tweet) is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. The training process involves assigning each word in each document to a random topic and then iteratively updating these assignments to maximize the coherence of words within topics while ensuring the documents reflect the mixture of topics found in the corpus. The strengths of LDA are that it is very scalable and easily interpretable, as well as being an easily implementable topic modelling algorithm. The weakness of LDA is that it does not account for homonyms or synonyms and using a BoW potentially loses important semantic information as it ignores syntax and order of words. Moreover, LDA generally performs comparatively weakly on smaller sample sizes (like our dataset) and smaller documents (like Tweets). We will evaluate our model based on coherence score and human judgement. The coherence score evaluates the quality of the learned topics by measuring the

degree of semantic similarity between high-scoring words in each topic. While coherence scores provide a quantitative measure of topic quality, human interpretation is crucial for assessing the relevance and interpretability of the identified topics. We reviewed the top words and representative tweets for each topic generated by the model to ensure they formed coherent and distinct themes that could be easily understood and labelled. It is also needed to be able to form the topics given the keywords per topic that the LDA model outputs. For example, if the LDA model identified a topic heavily featuring words like "Leafs," "hockey," "game," and "win," the coherence score could help confirm that these words frequently appear together in a contextually meaningful way across the dataset. Human judgement would then assess whether the topic accurately represents discussions around Toronto Maple Leafs hockey games, adding a layer of validation to the coherence score's indication of topic quality. There was no direct baseline model for our specific question, but we compared our coherence scores against scores from other LDA models with different numbers of topics to determine the optimal configuration.

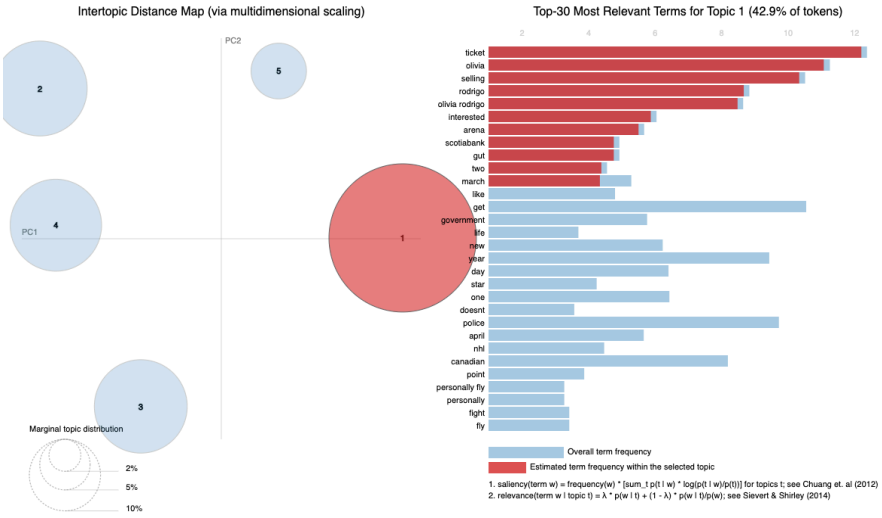
V. Results and Conclusions

We applied hyperparameter tuning to help decide the number of topics. We ran LDA with a different number of topics, computed the coherence score on our validation set, and then picked the topic number with the highest coherence score (see next page for graph).

Our model identified 4 discernable topics within the tweets and 1 topic that did not make sense. From the results, we determined that these 5 topics were the Olivia Rodrigo concert, the start of April, Canadian police, and Nathan Mackinnon's



goal against the Leafs. The last topic contained words that we determined to be incoherent. Additionally we can use the pyLDAvis library to help visualize our results:



We can see the distance between the topics as well as the predicted vs actual frequency of terms within each topic. The coherence score calculated for our model was 0.56 which suggests that, on average, the model has identified topics where the top words have a reasonable degree of semantic relatedness. This means the model has managed to capture themes that are somewhat interpretable and relevant to human understanding. We attested the issue of the incoherent topic to the fact that the nature of Twitter is such that the distribution of topics within the tweets we collected was too large and that there was no other prominent topic to group the rest of the tweets into. We realized for our LDA model to be successful, we have to have a “junk” topic where we can group the tweets that are not trending. Our findings provide a snapshot of the public discourse related to Toronto on Twitter, showcasing the diversity of interests and concerns among the city's residents and those discussing it online. Given more time and resources, expanding our dataset would likely improve our results substantially. Additionally, it could be enlightening to try different topic modelling algorithms like NMF and BERTopic and compare results.