JSC270 Winter 2024 Assignment 4: Natural Language Processing (NLP)

Christoffer Tan (1008740445) and Janis Joplin (1009715051)

Contributions Breakdown

This assignment was finished with collaborative effort, with both team members actively contributing to every aspect of the work. We worked jointly on each part of the code development, report, and presentation creation. We utilized Git to share code, Typst to write report, and PowerPoint to make presentation.

Part 1: Sentiment Analysis with a Twitter Dataset

A) Based on the training data, the proportions for each sentiment type are:

```
The proportion of the observations in the training data belong to sentiment type 0: 0.374122220872312
The proportion of the observations in the training data belong to sentiment type 1: 0.18738913862228163
The proportion of the observations in the training data belong to sentiment type 2: 0.4383914469687766
```

Note, in the result above, sentiment type 0 is negative, type 1 is neutral, and type 2 is positive.

B) We can tokenize the tweets by using the tokenizer 'punkt' from the nltk library. Some of the tokens for each tweet can be seen below.

	Label	Message	Sentiment	Tokens
0	0	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	1	[@, MeNyrbie, @, Phil_Gahan, @, Chrisitv, http...
1	1	advice Talk to your neighbours family to excha...	2	[advice, Talk, to, your, neighbours, family, t...
2	2	Coronavirus Australia: Woolworths to give elde...	2	[Coronavirus, Australia, :, Woolworths, to, gi...
3	3	My food stock is not the only one which is emp...	2	[My, food, stock, is, not, the, only, one, whi...
4	4	Me, ready to go at supermarket during the #COV...	0	[Me, ,, ready, to, go, at, supermarket, during...
...	...	...	...	...
41150	41150	Airline pilots offering to stock supermarket s...	1	[Airline, pilots, offering, to, stock, superma...
41151	41151	Response to complaint not provided citing COVI...	0	[Response, to, complaint, not, provided, citin...
41152	41152	You know itÂ□s getting tough when @KameronWild...	2	[You, know, itÂ□s, getting, tough, when, @, Ka...
41153	41153	Is it wrong that the smell of hand sanitizer i...	1	[Is, it, wrong, that, the, smell, of, hand, sa...
41154	41154	@TartiiCat Well new/used Rift S are going for ...	0	[@, TartiiCat, Well, new/used, Rift, S, are, g...

C) The `remove_url` function iterates through each row of the input data's `Tokens` column, using a regular expression (`re` library) to delete all tokens that is an url.

D) The first function, `remove_punctuation`, iterates through each row of the input data's `'Tokens'` column, utilizing a regular expression to eliminate all punctuation and special characters from each token. Then, the `convert_to_lowercase` function traverses through the `'Tokens'` column, converting each token to lowercase.

E) In the `stemming_tokens` function, we use `PorterStemmer` to reduce its token to its stem. We keep the stem of each token in a new list called `stemmed_tokens` and then replace `data['Tokens']` with that list.

F) The `remove_stopwords` function iterates through each row of the input data's `'Tokens'` column and keep only words that are not stopwords.

G) The length of the vocabulary is 74225

H) Based on Naive Bayes model to the data, we got the accuracy of the training data and the test data as below:

```
Test accuracy with simple Naive Bayes on training data: 0.8200529756263517

Test accuracy with simple Naive Bayes on test data: 0.6695629278567667
```

Also, the 5 most probable words in each class with their counts are provided below:

```
Class 0 (Negative):
   s: 43440
   coronavirus: 12545
   covid19: 13065
   price: 39670
   food: 20290
Class 1 (Neutral):
   s: 43440
   coronavirus: 12545
   covid19: 13065
   store: 48023
   supermarket: 48602
Class 2 (Positive):
   s: 43440
   coronavirus: 12545
   covid19: 13065
   store: 48023
   supermarket: 48602
```

I) In this scenario, we think ROC curve may not be appropriate because it is primarily designed for binary classification tasks, while in this case the there are three sentiment types. Hence, alternative metrics and evaluation techniques might be needed for this scenario.

J) After doing the TF-IDF transformer, we got the accuracy of the training data and the test data as below:

```
Test accuracy with simple Naive Bayes on training data: 0.7245510437170422

Test accuracy with simple Naive Bayes on test data: 0.6332280147446024
```

Hence, the accuracy of using count vectors are better than the TF-IDF transformer.

K) Instead of using stemming, we use lemmatizing and we got the accuracy of the training data and the test data as below:

```
Test accuracy with simple Naive Bayes on training data: 0.8345848217540278

Test accuracy with simple Naive Bayes on test data: 0.6727224855186941
```

As shown above, the accuracy of using lemmitization is slightly better than the accuracy of using stemming in both data sets.

Bonus The Naive Bayes model is generative as it tries to model $Pr(w_1, ..., w_D \mid S)$ to infer $Pr(S \mid w_1, ..., w_D)$ where $w_1, ..., w_D$ are the tokens in the document and S. That is, first we estimate the probability of getting a certain token given that the sentiment is S (S is either 0, 1, or 2). Then,

we use the Bayes rule to infer the probability we need, that is, the probability that the sentiment is S given our tokens and pick the sentiment with highest probability.

Part II

Trending Topic Analysis of Twitter Data using Latent Dirichlet Allocation and ChatGPT

Problem Description and Motivation

The rapid growth of social media platforms like Twitter has transformed information sharing and trend analysis. However, this task is complex due to dynamic trends, vast data volumes, and the presence of noise in discussions. Our Twitter dataset, comprising a substantial number of tweets (10,000) created within a short timeframe (approximately 4 days), aims to tackle these challenges head-on.

Existing studies, such as "*Sentiment Analysis of Twitter Data*" (Bagheri and Islam, n.d) and "Classifying Twitter Topic-Using Social Network Analysis" (Himelboim et al., 2017) have explored various approaches like sentiment analysis and network analysis. In contrast, our analysis employs Latent Dirichlet Allocation (LDA), offering a distinct advantage by uncovering latent topics within tweet text. This approach provides a deeper and more nuanced understanding of trends compared to simplistic frequency-based methods, thereby enriching insights gained from trend analysis on Twitter.

Data Description

The dataset used in this analysis is a pre-collected Twitter data from March 23th to March 27th, 2023, consisting of 10,000 observations and seven features. Among these features, the 'tweet_text' field is of primary interest for this analysis, requiring several pre-processing steps including: expanding contractions, removing mentions and tags, tokenizing the tweet, removing

URL, punctuation, converting tokens to lowercase, removing stopwords, and lemmatizing tokens to their base forms.

One crucial limitation of the dataset is that the tweet messages are not complete and may be cut off, which can obscure the full context of the topics discussed. This can lead to incomplete insights and potential misinterpretation of trends, especially if crucial information is missing due to tweet truncation. Despite the limitations, a significant strength of the dataset is its ample sample size, consisting of 10,000 observations. This large volume of data provides a robust foundation for analysis and enhances the statistical validity of the findings.

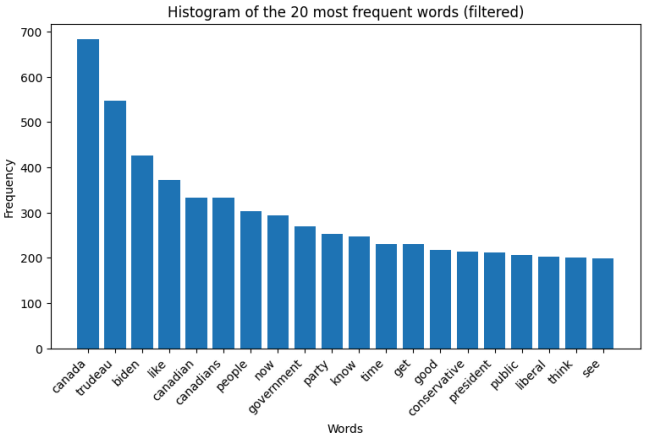
Exploratory Data Analysis

Before constructing our model, we conducted an exploration of the dataset's features. After careful consideration, we decided that only the 'tweet_text' field would be used for our model.

```
1 df.head()
```

	tweet_text	tweet_favourite_count	tweet_created_at	tweet_retweet_count	user_statuses_count	user_screen_name	user_followers_count
0	The priority for the city should be on providi...	0	2023-03-27 21:01:36+00:00	0	10194	EmergencyAgent	1145
1	Like it or not, this kind of clip is the sort ...	0	2023-03-27 21:01:26+00:00	0	2843	EDenhoff	4630
2	On a scale of moderately conservative (1) to r...	0	2023-03-27 21:01:16+00:00	0	13608	dzoolander85	11234
3	Keep up with the latest politics with The Dail...	0	2023-03-27 21:01:06+00:00	0	3610	VassKapelosShow	4901
4	So much resemblance to NAZI Germany leading up...	0	2023-03-27 21:01:05+00:00	0	108226	marshiehilgs	515

Following preprocessing steps outlined previously, we proceeded to analyze the dataset using visualizations such as word clouds and histograms.



The word cloud and histogram allowed us to observe the distribution of words within our dataset. Notably, we observed a significant number of tweets related to politics and government around Canada. This observation may indicate the occurrence of notable events or developments near the date of our dataset.

Machine Learning Model

For our analysis, we are utilizing Latent Dirichlet Allocation (LDA) as our machine learning model for clustering. LDA is an unsupervised clustering model used for topic modeling in text data, such as tweets. It assumes that each document is a mixture of topics, and each word in the document is generated by one of these topics. The model infers the distribution of topics across documents and the distribution of words across topics by iteratively assigning words to topics and adjusting the assignments to optimize the likelihood of the observed data. This process reveals the underlying topics in the dataset, allowing for a deeper understanding of the content and facilitating more nuanced analysis.

The strength of LDA lies in its ability to uncover latent topics within a large corpus of text data, enabling the identification of underlying topics. It allows for the discovery of complex patterns and relationships in the tweets. However, LDA has limitations such as the need for pre-defined topics or the potential for topics to overlap, which can lead to ambiguous interpretations. Additionally, LDA may struggle with short documents or noisy data, impacting the accuracy and the reliability of the topic modeling results.

To identify the optimal number of topics for our LDA model, we use metrics such as topic coherence, which measures the semantic similarity between words within topics. To facilitate this analysis, we split our initial dataset into three parts: the training (60%), validating (20%), and testing (20%) dataset. Using the validate dataset, we compare the coherence scores of different models and identify the number of topics that best captures the underlying topics in our tweet data.

After identifying the optimal number of topics (13), we construct our model using the *models.ldamodel* provided by *Gensim* with appropriate hyperparameters. The LDA model then generates 13 unlabeled topics, each accompanied by relevant words and their corresponding scores. We then input these words and scores into ChatGPT to assign a label to each topic. Finally, with this model, we can cluster the test data into topics that have the highest probability based on the output of our model and assign the corresponding label obtained from ChatGPT earlier.

Conclusion

Our model, built using LDA by Gensim and ChatGPT, efficiently assigns appropriate topics to tweet messages. Its rapid topic generation, often completed under one second, renders it a valuable tool for labeling datasets in supervised machine learning and enhancing semantic understanding in clustering algorithms. To evaluate our model's performance, we employ the same topic coherence metrics as described earlier. Upon running the topic coherence test on the testing dataset, the model achieved a score of 0.576. In comparison, our model performs similarly well to the baseline LDA model by iCAS Data Science & Analytics Forum (2023)

which also has the topic coherence scores ranging from 0.5 to 0.59. Given the high score attained on a dataset not used for training, we are confident that our model does not overfit, particularly since we utilize default values for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, known to avoid overfitting.

However, the model faces limitations. It occasionally assigns irrelevant topics to short tweets and struggles with tweets significantly different from the training data. Moreover, the absence of bigram interpretation may result in the loss of nuanced meaning from phrases. Despite these challenges, our model represents a notable advancement in topic analysis, offering efficient and insightful processing of tweet data. With further refinement, it holds promise for unlocking deeper insights from social media content and advancing analytical capabilities.

References

Bagheri, H. and Islam, Md (n. d). Sentiment analysis of twitter data. Iowa State University, United States of America

iCAS Data Science & Analytics Forum (2023). Latent Dirichlet Allocation (LDA) Topic Modelling in Python. Retrieved from https://www.casact.org/sites/default/files/2023-04/iCAS-4_Latent_Dirichlet_Allocation_Topic_Modeling_in_Python.pdf.

Kapadia, S. (2019). Topic Modeling in Python: Latent Dirichlet Allocation (LDA). Retrieved from https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0.

Li, S. (2018). Topic Modeling and Latent Dirichlet Allocation (LDA) in Python. Retrieved from https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24.