CSC401 A1 Part 4

Qien Song

February 2021

Introduction

There are many different approaches to classify the political bias inherent in a tweet. While part 3 uses a direct approach that fits the features on the labels, part 4 explores an alternative approach by identifying the tweets' topic through non-negative matrix factorization (NMF), and then training the

Methodology

1. Preprocessing
Extract verbs, adjectives, nouns with a character length ≥ 5 with the basic assumption that longer words tend to carry more meaning. These three types of words form the backbone of any meaningful sentence.
2. Convert Words using term frequency–inverse document frequency statistics
Fitting the training dataset with a TFIDF vectorizer.
3. Create a new feature data set
Fitting the TFIDF-fitted dataset using NMF to generate latent topics. The i^{th} row and the j^{th} column of the new dataset is the probability that the i^{th} tweet belongs to j^{th} topic.
4. Training the new feature dataset on the 5 classifiers from Part 3.1 and compare results

How to Run

python bonus.py -o $output\_folder$ -i $input\_json$

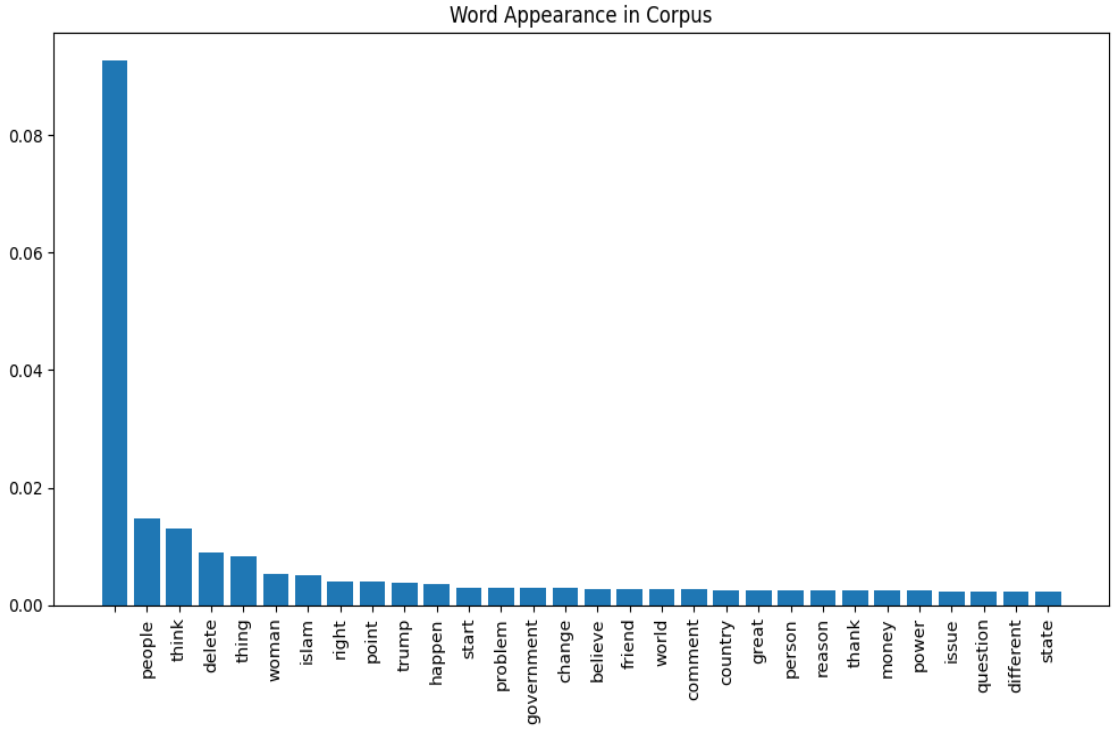
Result

The model's performance is around 5% worse comparing with 3.1. (Results are stored in the bonus folder with the same format as 3.1 output). There is also a higher level of skewness in certain classifiers' predictions, as they typically concentrate in one label. This does not necessarily mean that the topic-based

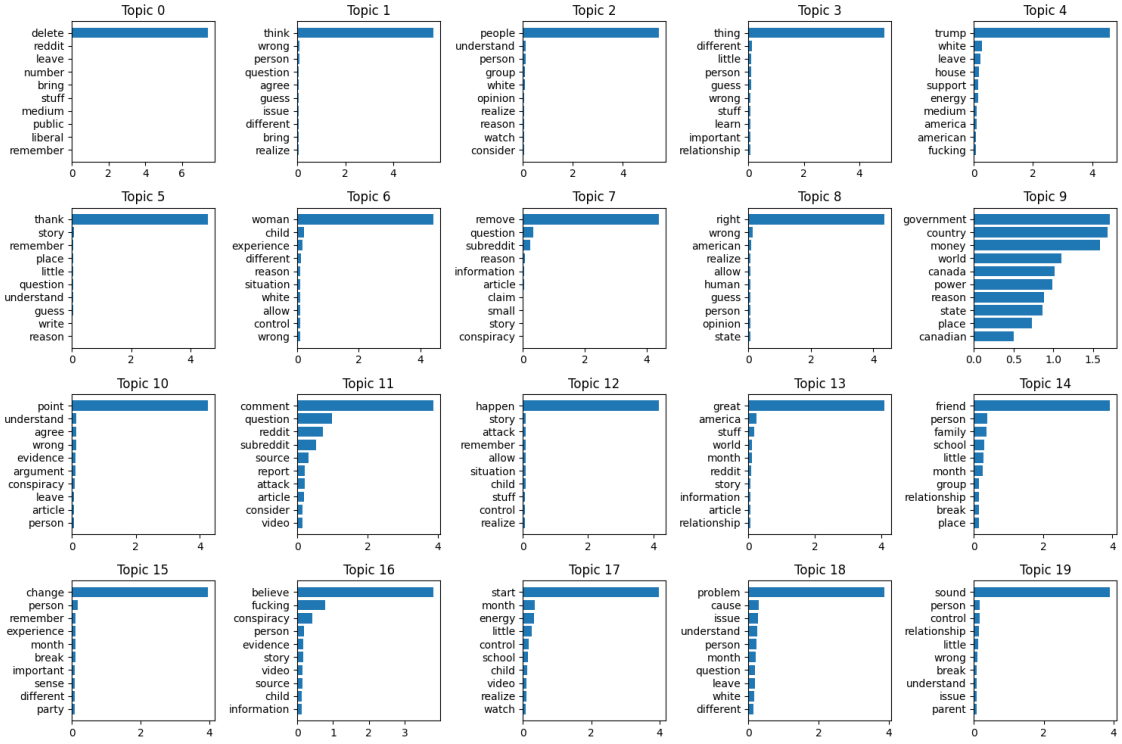
approach is an invalid one. In fact, it shows that topic, as a generic label, still holds decent explanatory power. If combined with other features, the incremental improvements may be more visible.

Analysis

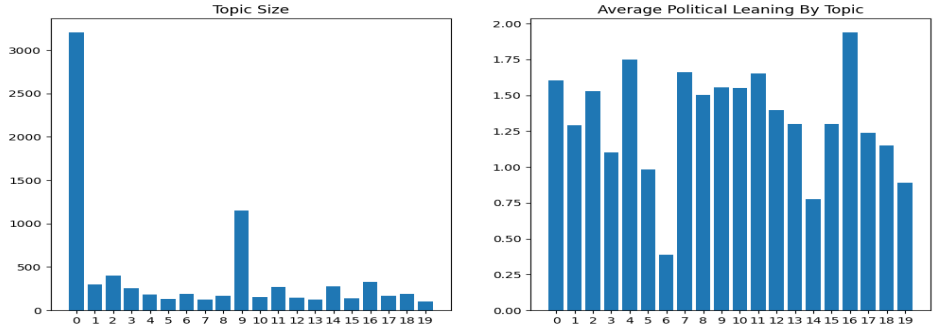
Examine the top words that appear in the training corpus, we see that most frequent words include empty strings (as there are no keywords inside the tweet), people, think, delete (signaling deleted tweets). Words that contain political connotation include "trump", "woman", "islam", "government" also appear.



Visualize the 20 topics that are created by NMF, we can make educated guesses on the political leaning of certain topics.



For example, Topic 4 features "Trump" as the top word, with keywords such as "white", "energy", "medium" (media), "america", may indicate that the tweets are likely to be 'Right' or 'Alt'.
On the other hand, Topic 6 that features "woman" and "child" could be more left-leaning.



To validate our hypothesis, we can examine the average political leaning of each topic. The most right leaning ones are Topic 16 and Topic 4. The most left leaning ones are Topic 6 and Topic 14, which matches our hypothesis.

Conclusion

In conclusion, using an unsupervised learning algorithm to supplement a supervised training task can reveal a lot of hidden patterns that exist in the corpus. With better integration between the two, the accuracy can foreseeably be improved.

Citation

https://shravan-kuchkula.github.io/topic-modeling/vectorize-the-reviews
https://towardsdatascience.com/unsupervised-nlp-topic-models-as-a-supervised-learning-input-cf8ee9e5cf28