Full name and semester: Saranjeet Singh Saluja, Spring 2022.

Title: Topic Modelling on Tweets related to Cancel Culture (1435 words)

Introduction:

Cancel Culture is a phenomenon in which people, brands or shows are 'cancelled' due to what some people feel is offensive behaviour or problematic ideologies. This subject has been a polarizing topic of debate over recent years. While many feel that it is a way of calling out for accountability for people's actions, many argue that it has devolved into a kind of social media mob rule. There have been studies to understand the process of 'cancelling' [1], to identify cancel culture as a form of social activism [2], to understand its impact on various industries [3] [4], and to identify potential factors (such as race, ethnicity, among others) [5] which could potentially influence the cancellation of a person. Through the study [6], the effects of 'cancel culture' on the ostracism of contrarians and the downfall of intellectual discussions within the domain of political science have been studied. This proposed study aims to contribute to this body of work by identifying the topics in a corpus of tweets from two groups, i.e., one which supports the idea of 'cancelling' and the other which is against the idea of 'cancel culture'.

Research Questions:

This study addresses the following questions:

- What are the most talked about topics in the tweets related to Cancel Culture?

Methods:

a. **Data Collection:**
The data for this study was collected from Twitter. The tweets were collected over the years from 2019 to 2021. The hashtags relevant to the subject – 'cancel culture', i.e., #cancelculture (the hashtag generally associated with the group opposing the very same idea), #cancelled and #yourecancelled (the hashtag generally associated with the group in favour of cancelling) were used to download the tweets from Twitter. For the same, two different datasets were created, each for pro and anti-cancelling.
The library 'snscrape' was used to collect the data. This library bypasses Twitter's restrictions and rate limits. Additionally, one does not require a Twitter developer account when one uses 'snscrape'. As I had to create a large dataset, I felt that 'snscrape' is the ideal choice for scraping the data.

b. **Data Cleaning:**
After scraping the data, I had 126988 data points for the pro-cancellation dataset, whereas 267376 data points were generated for the anti-cancellation dataset. To clean the data further, the posts containing less than three words, only hashtags and/or emojis were filtered out using regex to avoid reducing the quality of the analysis. As we are trying to identify abstract topics, Emojis will not contribute to the same. The tweets which were not in English, along with duplicate tweets were also filtered out.

Furthermore, lemmatization was performed using the WordNetLemmatizer from the nltk library. Based on vocabulary and morphological analysis, lemmatization reduces the word to its root form. Furthermore, any words beginning with symbols such as '', '&amp', and so on were excluded from the tweets. Finally, any mentions and punctuation in the tweets were removed.

c. Analysis:

For the study, sklearn's Latent Dirichlet Allocation algorithm was used to carry out Topic Modelling on the tweets. Latent Dirichlet Allocation algorithm uses probabilistic text analysis methods to detect abstract topics from a corpus of texts. While other libraries, such as 'gensim,' implement the same algorithm, I chose to use sklearn's version due to familiarity with the library and its effectiveness. The entire process was carried out in the following steps:

i. Loading the data:

The CSV file containing the tweets was loaded as dataframe using the pandas library. There were 187066 tweets for the anti-cancellation dataset, while 82297 tweets were present for the pro-cancellation dataset. The data frame to be worked upon consists of 'Datetime', 'Tweet Id', 'Text', 'Username', 'Language' and 'CleanedText' as shown in Fig.1 and Fig.2.

	Datetime	Tweet Id	Text	Username	Language	CleanedText
0	2021-12-31 23:44:41+00:00	1477063190349549573	would she still be an elitist joshua if she le...	CharlieRClaywel	en	would still elitist joshua leave nyt bubble mo...
1	2021-12-31 23:25:45+00:00	1477058426261762050	typical liberal and bully tactic should backfi...	voiceoreason702	en	typical liberal bully tactic backfire circle w...
2	2021-12-31 22:40:00+00:00	1477046909248360453	how have or not take issue with the	FreemanSprings	en	take issue
3	2021-12-31 22:19:38+00:00	1477041786216349697	not a pier morgan fan but people clutch at str...	sokyola87	en	pier morgan fan people clutch straw chat someo...
4	2021-12-31 21:57:22+00:00	1477036182059978754	the prejudice toward intelligence continue to...	BobBOSS92280105	en	prejudice toward intelligence continue oppress...

Fig.1. Data Frame of the Anti-Cancellation Tweets

	Datetime	Tweet Id	Text	Username	Language	CleanedText
0	2021-12-31 23:04:42+00:00	1477053128402952199	im not a girl why do you hate men j	akisbarchie	en	im girl hate men j
1	2021-12-31 22:56:24+00:00	1477051036904865800	if you have a flight to denver airport you may...	K__TequilaShots	en	flight denver airport may want check airline s...
2	2021-12-31 22:33:56+00:00	1477045384518291462	my mate mum meet jodie im star buck a few week...	Sapphicmarla	en	mate mum meet jodie im star buck week ago deci...
3	2021-12-31 22:17:21+00:00	1477041210376302599	cancel pirate yet again for exist have a nice ...	elysiummods	en	cancel pirate yet exist nice new year
4	2021-12-31 22:08:40+00:00	1477039028004392963	and cup have be know to like the ugly men i v...	roadtwist	en	cup know like ugly men ever see life tbh doubl...

Fig.2. Data Frame of the Pro-Cancellation Tweets

ii. Converting the Text to Matrix and Finding appropriate Number of Topics:

Using sklearn's CountVectorizer, a sparse matrix of the words along with their count per document was created which would be further fed to the LDA algorithm as input. The cleaned text for both datasets was converted to count matrices. Selection of the appropriate number of topics is necessary as it allows to identify the distinct number of topics and the words associated with those topics. For the same, sklearn's RandomizedSearchCV was used to identify the appropriate

number of topics, along with other hyper-parameters such as learning rate and max iterations. This hyper-parameter tuning method uses random combinations to find the best hyperparameters. This method was carried out for both datasets, yielding the hyperparameters shown in Fig. 3 and Fig. 4.

```
Best Params of LDA Model for Cancel Culture Dataset:  {'n_components': 3, 'max_iter': 30, 'learning_decay': 0.5}
Best Score of LDA Model for Cancel Culture Dataset:  -2716054.092893912
```

Fig.3. Best Hyperparameters for Anti-Cancelling Dataset

```
Best Params of LDA Model for Cancelled Dataset:  {'n_components': 3, 'max_iter': 20, 'learning_decay': 0.7}
Best Score of LDA Model for Cancelled Dataset:  -792124.8943824958
```

Fig.4. Best Hyperparameters for Pro-Cancelling Dataset

From Fig. 3 and Fig. 4, it is apparent from the hyperparameter – 'n_components', that the appropriate number of topics for both the datasets is three.

iii. Plotting graphs to interpret results:

An essential component of understanding the type of topics during the established timeline is graph plotting. The topics for each dataset were plotted with the help of a word cloud, using the WordCloud library. A word illustrates the words associated with the topic based on how frequently the word appears within that topic.

Results:

a. Visualization using Word Cloud:

A. Anti-Cancelling Dataset:

- Topic 1:

  Fig. 5, illustrates the words associated with Topic 1. We can observe the words new, come, talk, today as the most predominant words. Additionally, we have words such as podcasts, video, history, read, trump and listen which may indicate promotions of such mediums centred around the subject of cancel culture.

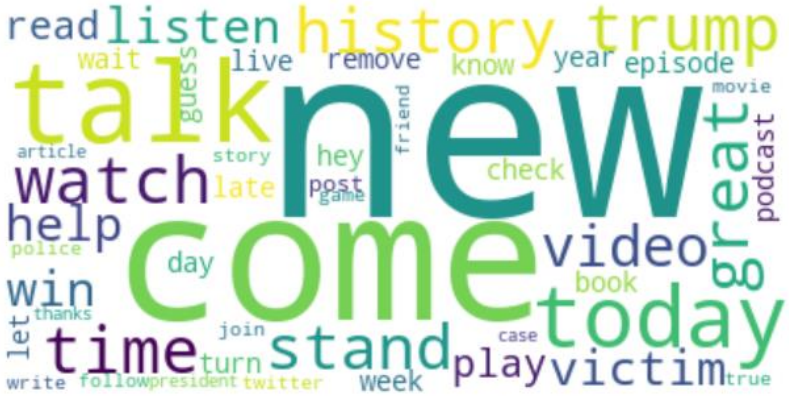

Fig.5. Word Cloud Visualization for Topic 1 of Anti-Cancellation Dataset

- Topic 2:
  Fig. 6, illustrates the word cloud visualization of Topic 2. The most frequent words associated with this visualization are cancel, culture, people, don't, think and make. From these words, it is safe to assume that Topic 2 talks about cancel culture and how people do not think before cancelling someone.

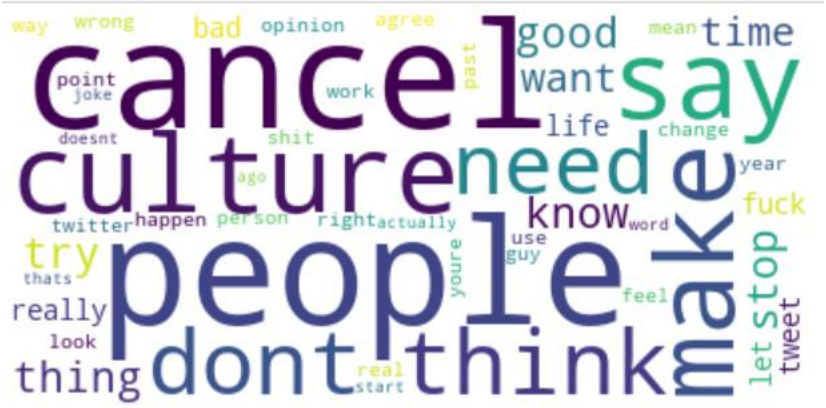

Fig.6. Word Cloud Visualization for Topic 2 of Anti-Cancellation Dataset

- Topic 3:
  Fig. 7, denotes the most frequent words linked with Topic 3. In this topic, the prevailing words are racist, left, right, medium and mob. This topic most likely is about the political nature of cancel culture i.e., people inclined towards anti-cancelling are more likely to be right or far-right aligning.

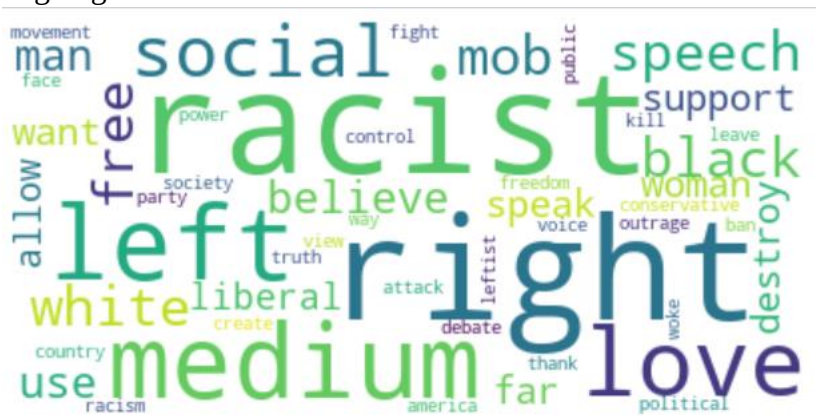

Fig.7. Word Cloud Visualization for Topic 3 of Anti-Cancellation Dataset

B. Pro-Cancelling Dataset:

- Topic 1:

The words new, cancel, year, season, service, and flight depict the most frequent words within this topic. From the words based on the word cloud visualization of Topic 1, it appears that the subject of Topic 1 is primarily cancelling a service (either a TV show, movie, event, flight or even year).

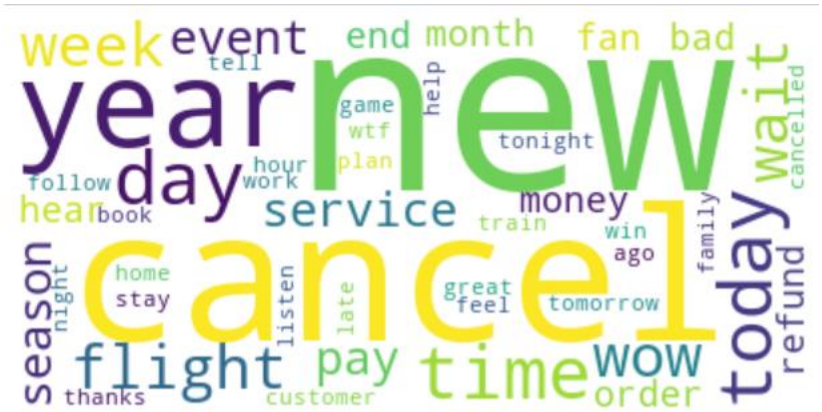

Fig.8. Word Cloud Visualization for Topic 1 of Pro-Cancellation Dataset

- Topic 2:
  Based on the words in the word cloud visualization, the most appropriate subject for Topic 2 would be cancelling someone (from the words cancel, and believe) or how cancelling is required (from the words need, say, good).

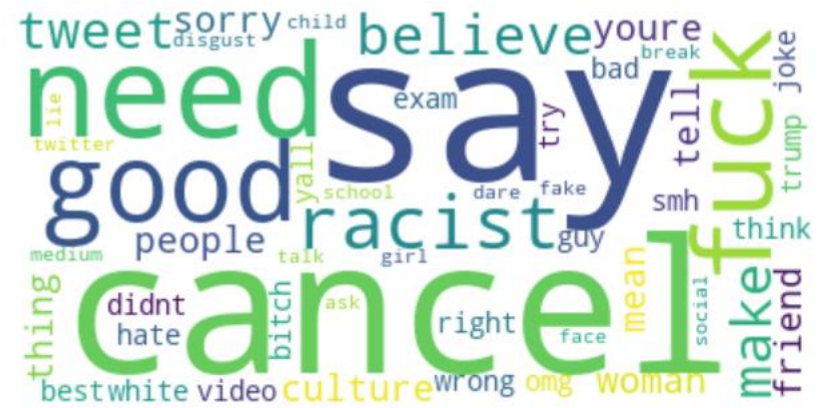

Fig.9. Word Cloud Visualization for Topic 2 of Pro-Cancellation Dataset

- Topic 3:
  Fig. 10, illustrates the word cloud visualization of Topic 3 of the Pro-Cancellation dataset. The predominant words within this topic are people, know, don't, think, look, and come. The topic does not seem to have a distinct subject, but it could be said that this topic represents the circumstances for cancelling (from the words support, officially, person).

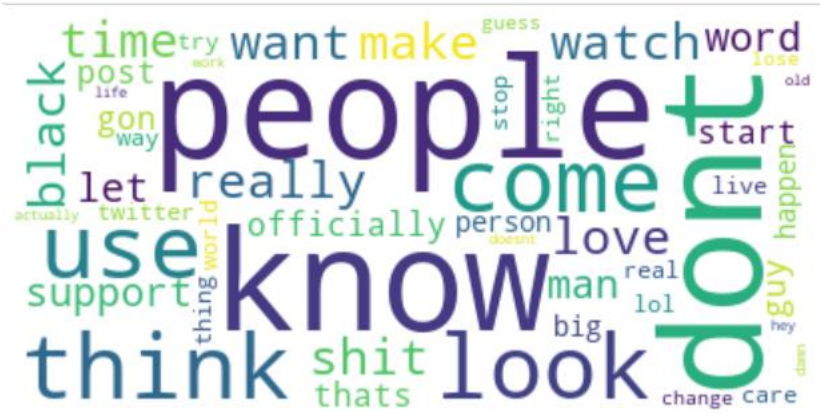

Fig.10. Word Cloud Visualization for Topic 3 of Pro-Cancellation Dataset

Conclusion:

In this research, topic modelling was performed on a corpus of tweets from two groups, i.e., one which supports the idea of 'cancelling' and the other which is against the idea of 'cancel culture'. The idea behind this was to get an insight into the topics which are discussed by people having different perspectives on the specific action of 'cancelling' someone. While we do get clear and distinct topics for people against 'cancel culture', there is no such distinction for Topic 3 for people in favour of 'cancelling' someone. However, Topic 1 and Topic 2 of the Pro-Cancellation dataset, give us a fair idea about potential subjects discussed by users in their tweets.

Limitations:

One of the main limitations of this study is the presence of noise in the form of irrelevant posts. For example, tweets about the cancellation of flights are not related to the concept of cancel culture but are still present as a major component of Topic 1 as observed in Fig. 8. Additionally, the cause of the lack of distinction in the topics in the pro-cancellation dataset could be attributed to the lack of tweets and/or length of the tweets as compared to the anti-cancellation dataset.[6]

References:

[1] Samantha Haskell, "Cancel Culture: A Qualitative Analysis of the Social Media Practice of Cancelling", Boise State University, 2021.

[2] Korri E. Palmer, "KancelKulture: An Analysis of Cancel Culture and Social Media Activism Through the lens of Minority College Students", The College of Wooster, 2020.

[3] Martinez, Alix, "Uncovering the Dirt on Cancel Culture: An In-depth Analysis of Publishing's Relationship with Controversy" (2021). Book Publishing Final Research Paper. 58.

[4] Bulgakova Alina Andreevna, Manich Anna Aleksandrovna, and Fomina Natalia Denisovna, "The Impact of Cancel Culture On Business." (2021). IX International Scientific and Practical Conference, Penza, 2021.

[5] Burmah, Loydie Solange, "THE CURIOUS CASES OF CANCEL CULTURE" (2021). Electronic Theses, Projects, and Dissertations. 1289.

[6] Pippa Norris, “Closed minds? Is a ‘cancel culture’ stifling academic freedom and intellectual debate in political science”, Harvard University, 2020.

[7] Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, Ming Zhang, “Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis”, Proceedings of the 31st International Conference on Machine Learning, PMLR 32(1):190-198, 2014.