AMSI VACATIONRESEARCH SCHOLARSHIPS 2021–22

Get a taste for Research this Summer

A Large-Scale Text Analysis of Australian TV Media

Tyson Rowe
Supervised by Professor Lewis Mitchell
University of Adelaide

Vacation Research Scholarships are funded jointly by the Department of Education and the Australian Mathematical Sciences Institute.

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Abstract

This project will develop and use natural language processing techniques to perform a text analysis on the Australian free-to-air commercial television media networks, Seven, Nine and Ten, using a unique data set: a collection of teletext closed captions provided by the tveeder service. The data set per channel comprises of hourly increments containing the closed captions from the 1st of January 2021, to the 12th of December 2021.

We started by analysing the information entropy and the distributions of words in each of the channels per hour, then focused on topic modelling with a refined version of the larger data sets, which consisted of only the 4-8 PM time period per day. This was done to explore the key topics in the news on these channels throughout the year, such as the 2021 covid lockdowns.

Contents

1 Introduction

Since the 20th century Australian TV media has become an integral part of modern culture. For an individual the Australian TV media landscape has a number of functions; it provides an accessible source of entertainment, whether it be through game shows, movies or sports broadcasts, it provides a means to garner knowledge with documentaries and it provides a way for us to keep up to date with the most recent news. In particular, with the rise of the coronavirus (covid) in early 2020, the importance of the news representation of covid in Australian TV media has come to the forefront. It is critical that covid is represented in an easily digestible and informative manner to help contain the pandemic, spread awareness and protect the mental health of individuals, especially with increasing misinformation about the pandemic spreading online by social media (Mach et al., 2021) (Su et al., 2021).

With the introduction of the BROADCASTING SERVICES ACT in 1992, broadcasters in Australia are required to caption all programs throughout the day to allow for deaf individuals to view their broadcasts. Tveeder is a service that reads these captions for free-to-air television networks and uploads them for the general public to use (Zhou, 2021). We will use the captions stored by tveeder to construct three data sets, one each for the networks Seven, Nine and Ten, consisting of hourly increments from the 1st of January 2021, to the 12th of December 2021.

We start our report by cleaning each of these data sets, and then undergoing some general analysis. We will explore the distrubution of words in each of the data sets, as well as the information entropy throughout the daily cycle for the year, to determine if there are any similar broadcasting patterns. We then narrow down each of the data sets to a 4-8 PM time period to undergo topic modelling across each of the channels. This will help us explore the major topic of covid in the news throughout the year of 2021.

Statement of Authorship

The workload was divided as follows:

- Fellow University of Adelaide student, Ben Lang, collected all the teletext closed caption data using the tveeder API in Python and uploaded the captions to a shared drive.
- Tyson Rowe cleaned and analysed this data using Python and produced all results and interpretations wrote in this report.
- Professor Lewis Mitchell supervised the project, provided input on project direction, assisted with interpreting the results and proofread this report

2 The Data

The captions from the Australian TV media were collected from tveeder using the tveeder *API* in Python. An API (Application Programming Interface), is a programming interface that simplifies the interactions between different software. In this case, tveeder reads the captions provided by the television networks and stores them, then the programmer is able to retrieve any stored data using the tveeder API. As mentioned in the introduction, our focus will only be on the free to air commercial networks Seven, Nine and Ten.

The data set per channel initially comprised of 8314 hourly increments of teletext closed captions between the start of January 1st 2021, to the end of December 12th 2021. Each hourly increment was saved as a txt file, and labeled uniquely with the date and hour of day for the given increment. We then removed any txt files that had no text inside them. This resulted in the following new sizes for each of the data sets (in hourly increments); Seven has 8176, Nine has 7669, and Ten has 7344. The removed hourly increments mostly came from the 2-5 AM morning period.

2.1 Cleaning the data

As each of the data sets for the channels are large and messy, we require to clean the data to a more easily accessible form before undergoing any analysis. To do the cleaning process, the Python package NLTK (Natural Language Tool Kit) was used. Starting with reading in a large string of closed captions for one hourly increment, the cleaning process can be broken down into the following four steps as follows:

1. We start by tokenising the string into *tokens*. A token is a smaller piece of the string, normally a word or a punctuation character.

2. We then remove any punctuation and set all characters to lower case in these tokens.

3. We then remove *stop words* from the tokens. A stop word is defined as a highly common word in the English language. Examples include; 'will' or 'be'.

4. We finally *lemmatise* the tokens. A lemma is the simplest form of a set of words. For example, the set of words, 'break', 'breaks', 'breaking' and 'broke' have the lemma 'break'.

By cleaning the data, in particular by removing any stop words, we remove redundant information that we know will appear plenty of times throughout the text. Furthermore, by removing uppercase, as well as lemmatising the tokens, we help to merge tokens that have the same meaning, which helps to remove any unnecessary skew in the data. This will be particularly important when we come to topic modelling.

2.2 Distrubution of words

We start our general analysis by understanding the different vocabulary distributions across the channels. To do this we have plotted the rank of a word against it's frequency for the year of data for each channel. This is shown in figure 1 with log-log (base ten) axes. The frequency of each word was determined using the counter package in Python.

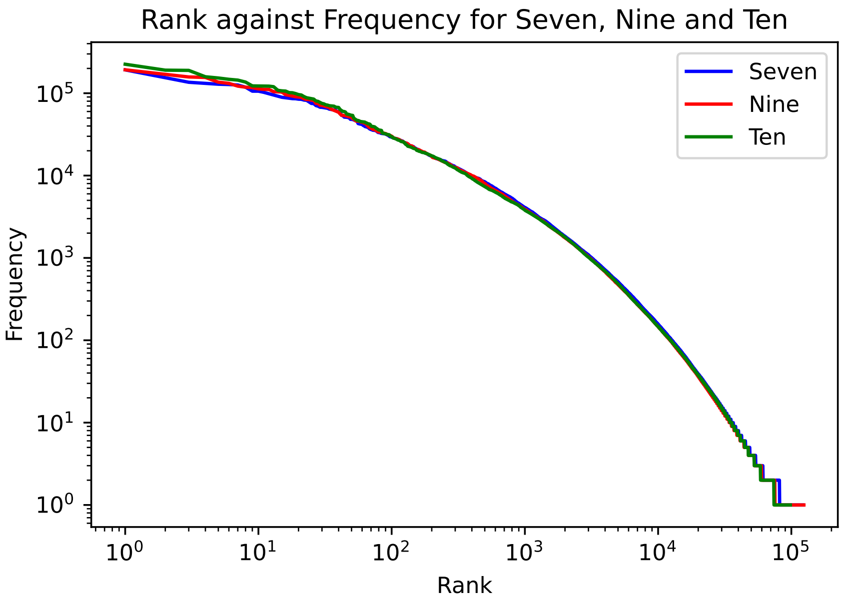

Figure 1: The rank of a word against frequency for each of the different cleaned data sets.

From figure 1, we can see that the rank-frequency plot is very similar for all the channels, especially for word rankings of 100 and onward. This is interesting, as it suggests similar vocabulary distributions, as well as may indicate a common underlying structure of each of the channels in terms of show programming throughout the year.

2.3 Information Entropy

To gain a deeper understanding of the vocabulary distributions throughout the diurnal cycle, we will use *information entropy*. Information entropy provides a measure of the underlying uncertainty in a probability distrubution. An entropy of zero means that there will be zero uncertainty of the probability distrubution, while a higher entropy implies more uncertainty. There are many different types of information entropy, but we have opted to use Shannon's entropy in log base 2 (with units of bits), given by the following equation (Shannon, 1948):

$$H(X) = -\sum_{i=1}^{N} p(i) \log_2 p(i),$$

where N is the total number of unique words in the text and $p(i)$ is the probability of the i^{th} word occurring in the text. In this instance, we define the text as the hourly increments across each of the channels. Computing the entropy for the sets of data, the mean entropy and variance for each of the channels is shown below in table 1.

Channel	Seven	Nine	Ten
Mean (bits)	9.41	9.37	9.37
Variance	0.41	0.46	0.71

Table 1: The mean entropy (in bits) and variance for the data set per hourly increment.

From table 1, we see that the entropies for all three channels are roughly the same, with Seven being the most uncertain, followed by Nine and Ten being equal. However, we see that the variance for Ten is much larger compared to Seven and Nine, suggesting there is more spread across the hourly increments. To explore this further, consider plotting the mean entropy per hour throughout the day for all three channels, as seen in figure 2.

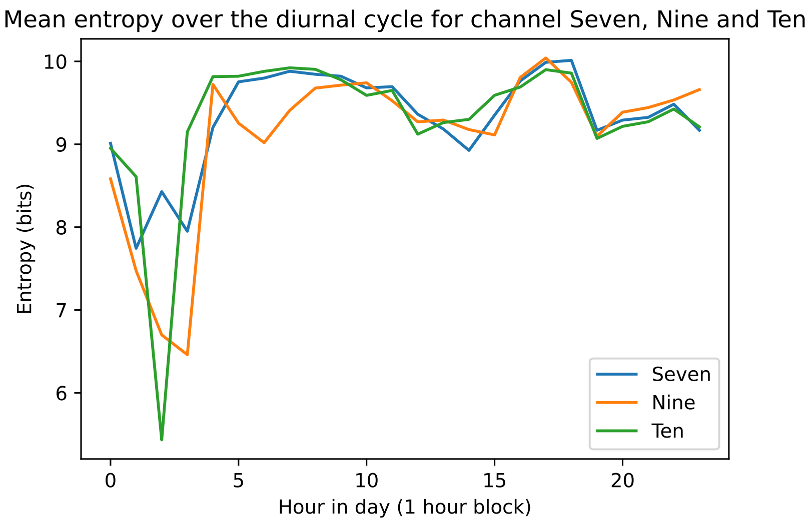

Figure 2: The mean entropy (in bits) of each of the channels over the diurnal cycle.

From figure 2, we see that diurnal cylce of the entropy is approximately the same across all three channels. This supports the claims made with figure 1, of a potential underlying common structure in the channels. Furthermore, we see that at 3 AM, the entropy of Ten drops a lot compared to the other channels, thus explaining the differences in variance observed in table 1.

We notice that the entropy of the morning period is less than the entropy of the afternoon period, with all three channels peaking at 6-8 PM. We expect this result as the news is on at this time, which covers a number of different topics each day throughout the year. As such, there is more uncertainty on what will be presented and thus a larger entropy.

3 Topic Modelling

We now want to explore the main themes and ideas across all the channels. We will do this using a natural language processing technique called topic modelling, which is an algorithm that finds underlying hidden *topics* in a *corpus*. A topic is defined as a distribution of words over a fixed vocabulary, while a corpus is defined as a collection of documents (or texts).

The topic modelling algorithm used for this report is *LDA* (Latent Dirichlet Allocation). LDA is a generative probabilistic model that uses Bayesian inference to determine the most likely topics in a set of documents. A basic explanation of LDA can be given as follows; the LDA model assumes some distrubution of topics over the documents and some distrubution of words over the topics (both Dirichlet distributions), then it updates these distributions for the given observed data, the words in the documents (Blei, 2012).

Further understanding of the LDA model is not required for this report. We will determine all of the topics using LDA Mallet in Python, which is accessible through the Gensim package.

3.1 Filtering and choosing the number of topics

Due to the limited scope of this project, we have undergone some filtering on the data to reduce the run time of computations, as the current data set is too large to topic model all of it. As such, we have opted to only use the hourly increments in an afternoon time period (defined as 4-8 PM) across all channels. This is the preferred time period as it contains all the afternoon news programs across the three channels for all days of the week.

Furthermore, as we are only interested in the key topics across this time period, we have done the following filtering to the words in the corpus;

- Any words in less than 10 documents are removed, and
- Any words in more than 85% of the documents are removed.

This filtering was done using the Gensim corpora library.

After filtering, we require to determine an appropriate number of topics. There are many ways to do this, but we will use a useful score called the *coherence score*. The coherence score is a measure based on the similarity of words in a topic, that gives an indication of how related words are across a topic. We will use the 'c_v' measure of the coherence score, which is based off of pointwise mutual information and cosine similarity (Kapadia, 2019). A higher coherence score is preferred, as it tells us the topics are more closely related.

We calculated the coherence score for all the channels with a number of different topics (k) between 1 and 100, using the Gensim library in Python. A plot of the coherence score against the number of topics (k) has been produced in figure 3.

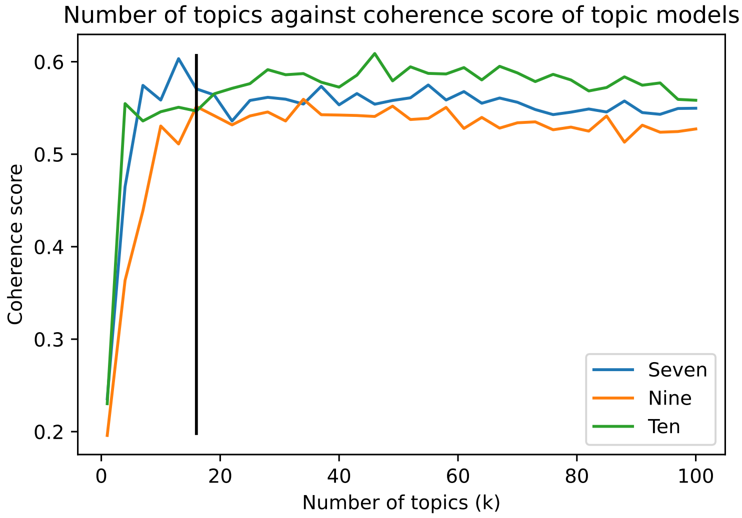

Figure 3: The coherence score of the topic models against the number of topics for all the channels. The chosen value of $k = 16$ has been indicated with the black line.

Using figure 3, we want to choose a number of topics that can be used for all channels. Ideally we want to use one of the peak coherence scores, but we will unnecessarily be adding topics if we decide on one of the different peaks, which occur at 13, 34 and 46 for Seven, Nine and Ten respectively. However, we notice that the coherence score flattens out at around 0.54, for k greater than 4, 7 and 16 for Seven, Nine and Ten respectively. As such, we choose $k = 16$ as the number of topics for all three channels.

3.2 Topics

We now present the results found from topic modelling. The most interesting topics found for channel Seven are shown below in table 2, and the daily proportion of time spent in these topics throughout the year is shown in figure 4. For a full list of the topics, please refer to the appendix. The 'likely topic' column was determined with the author's discretion.

Topic No.	Top 5 Common Words	Likely Topic
1	goal, ball, kick, player, ground	AFL
4	case, lockdown, covid, site, health	Covid lockdown
6	shower, rain, sunshine, afternoon, wind	Weather
9	ball, wicket, player, cricket, shot	Cricket
10	gold, medal, olympic, tokyo, aussie	Tokyo Olympics
11	police, covid, government, state, country	Covid / Government

Table 2: Most interesting topics derived from topic modelling on the filtered channel Seven data set. For the remaining topics, please refer to the appendix.

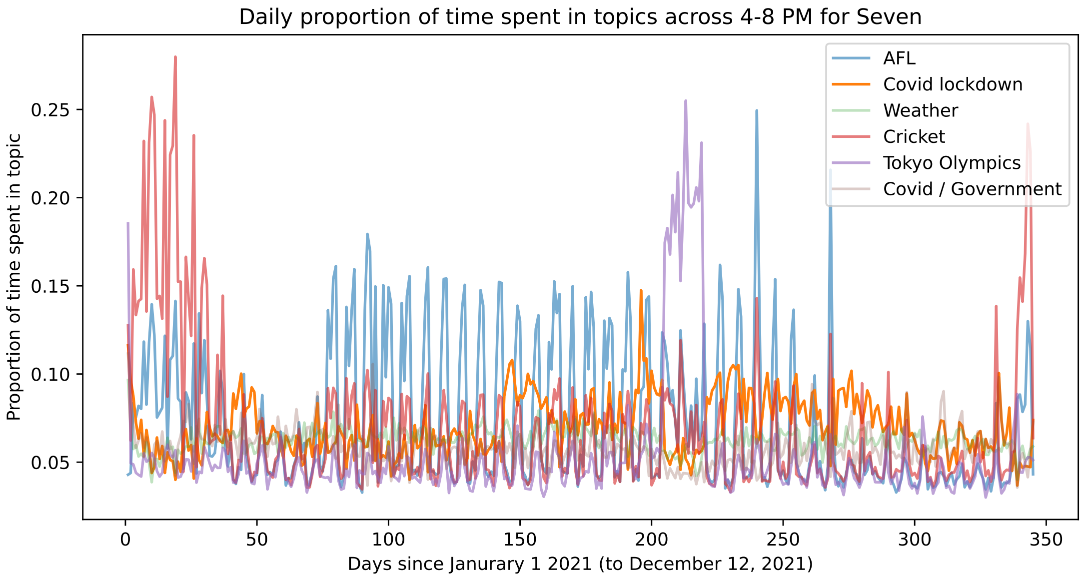

Figure 4: The daily proportion of the time spent in topics (listed in table 2) for Seven from January 1st 2021 to December 12th 2021.

From figure 4, we can see how the topic modelling has matched the sporting topics - cricket, AFL and Tokyo Olympics - with their respective airing times throughout the year. We have the cricket at the start of the year with the BBL season and test matches, followed by the AFL season starting early March and ending late September with the the Tokyo Olympics in between for two weeks starting late July. For the remaining topics, we see that they all have relatively stable proportions, with only the covid lockdown topic changing slightly throughout the year. We expect this result, as the weather will always occur in the news and both topics on covid were prevalent discussion points throughout the year of 2021.

The most interesting topics found for channel Nine are shown below in table 3, and the daily proportion of time spent in these topics throughout the year is shown in figure 5.

Topic No.	Top 5 Common Words	Likely Theme
1	police, christmas, covid, vaccination, child	General covid
9	shower, degree, afternoon, sydney, wind	Weather
11	room, block, space, bedroom, mark	The Block
12	match, court, player, serve, tennis	Tennis
16	lockdown, vaccine, covid, health, site	Covid lockdown

Table 3: Most interesting topics derived from topic modelling on the channel Nine data set. For the remaining topics, please refer to the appendix.

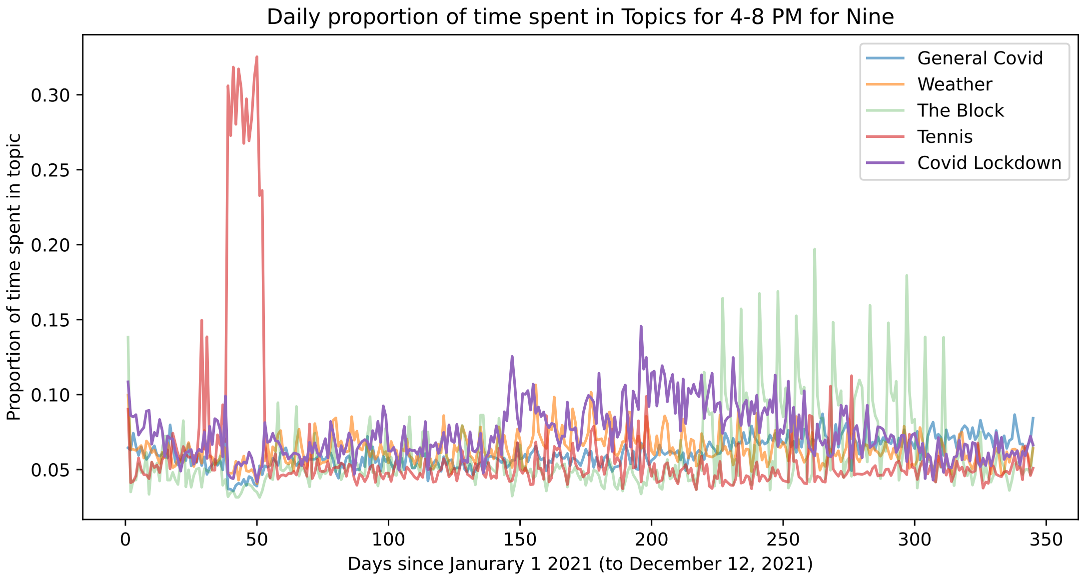

Figure 5: The daily proportion of the time spent in topics (listed in table 3) for Nine from January 1st 2021 to December 12th 2021.

From figure 5, much like in figure 4, we can see how the topic modelling has matched the sporting topic, Tennis, and the TV show topic, The Block, with their respective airing times throughout the year. We have the Australian open starting in February, and The Block in early August. We also notice some slight peaks in the Tennis topic throughout the year outside of the Australian open, which might coincide with the other Grand Slams being talked about in the News. For the remaining topics, they keep relatively stable proportions throughout the year, which again is the expected observation.

The most interesting topics found for channel Ten are shown below in table 4, and the daily proportion of time spent in these topics throughout the year is shown in figure 6.

Topic No.	Top 5 Common Words	Likely Theme
5	project, story, question, government, break	The Project
7	lockdown, sydney, state, heath, south	Covid lockdown
8	vaccine, government, quarantine, hotel, health	Vaccination
10	game, tomorrow, degree, shower, player	Sport / Weather
13	dish, cook, flavour, beautiful, mate	Master Chef

Table 4: Most interesting topics derived from topic modelling on the channel Ten data set. For the remaining topics, please refer to the appendix.

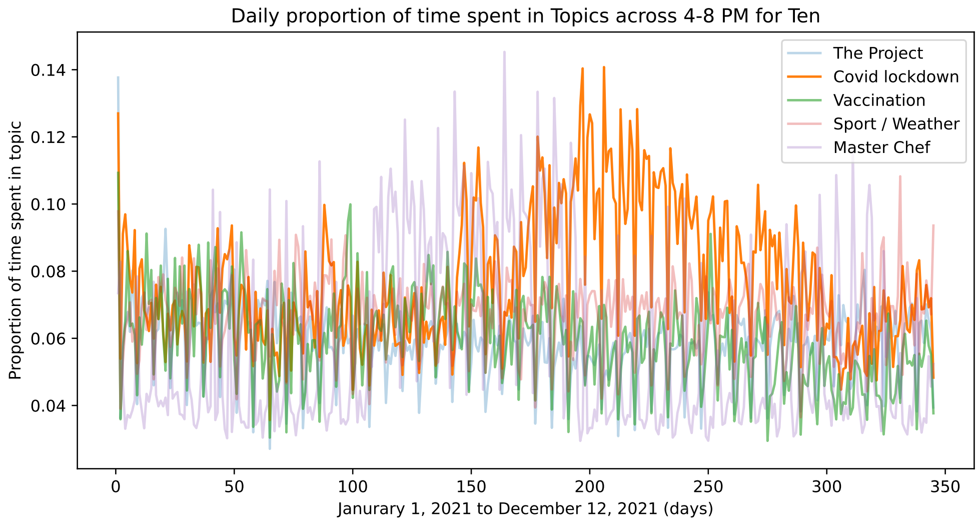

Figure 6: The daily proportion of the time spent in topics (listed in table 4) for Ten from January 1st 2021 to December 12th 2021.

From figure 6, we see that all of the topics, other than Master Chef which starts in April, appear to have appear to have stable proportions throughout the year. This is expected as they all feature in regularly scheduled TV programs. One interesting observation is the stable presence of the vaccination topic as the year progressed, which we would expect to see an increase. This might be due to the fact that the vaccination topic does not only focus on vaccination, as evident by the other key words in table 4.

3.3 Topic comparisons

We now want to compare the topics across the channels. There are many ways to do this, but we have opted to use the *JSD* (Jensen-Shannon Divergence). The JSD gives a symmetric measure of how similar two probability distributions, P and Q, are from an average distrubution, M,

$$M = \frac{1}{2}\Big(P + Q\Big).$$

In this instance, we think of P as the topic in one channel and Q as the topic in another channel. As P and Q have different vocabularies, we denote the combined vocabulary for M as X. Using the combined vocabulary, if $x \in X$ is a word, then the JSD is defined as (Lin, 1991),

$$JSD(P \,||\, Q) = \frac{1}{2} \sum_{x \in X} \left(P(x) \log_2 \frac{P(x)}{M(x)} + Q(x) \log_2 \frac{Q(x)}{M(x)} \right). \tag{1}$$

For (1), we require to define $0 \log_2 0 = 0$ by using the limit as $x \to 0$ of $x \log_2 x$.

We interpret the JSD as follows: a JSD of 0 implies the topics are the identical, while a JSD of 1 indicates that the topics are completely different. Now, computing the topic comparisons for Seven against Nine, we get the following coloured plot in figure 7.

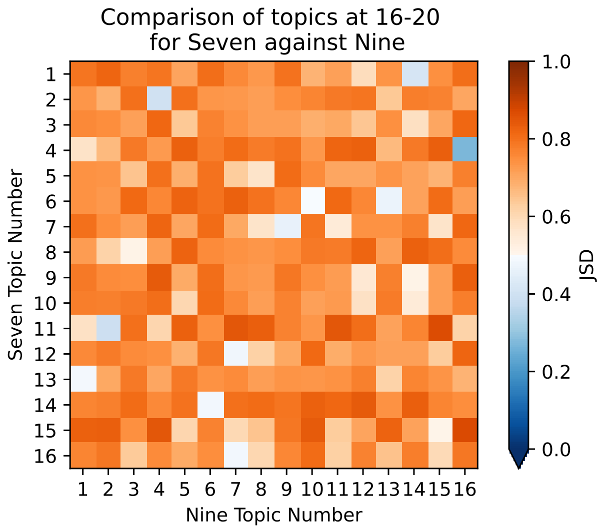

Figure 7: Topic comparisons of Seven against Nine for the 4-8 PM time period for 2021.

From figure 7, we observe that majority of the topics across Seven and Nine are sufficiently different, as evident by orange being the predominant colour suggesting JSD greater than 0.5. However, some notably similar topics are; 11 (Coivd/Government) against 2 (Unknown), 2 (Royals) against 4 (Royals) and 4 (Lockdown) against 16 (Lockdown).

Next, the topic comparisons for Seven against Ten are shown in figure 8.

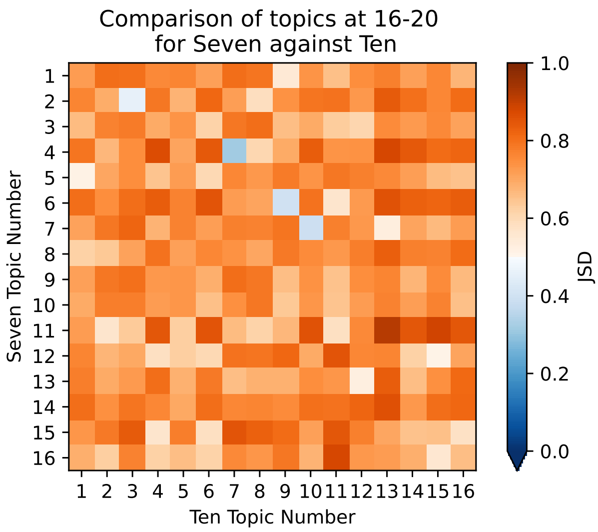

Figure 8: Topic comparisons of Seven against Ten for the 4-8 PM time period for 2021.

From figure 8, we again observe that majority of the topics across Seven and Ten are sufficiently different, as evident by orange being the predominant colour suggesting JSD greater than 0.5. However, some notably similar topics are; 4 (Lockdown) against 7 (Lockdown), 6 (Weather) against 9 (Sport / Weather) and 7 (Gardening) against 10 (Gardening).

Finally, the topic comparisons for Nine against Ten are shown in figure 9.

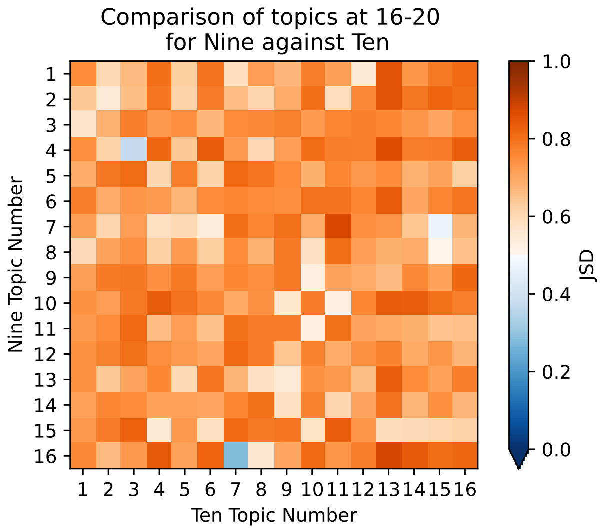

Figure 9: Topic comparisons of Nine against Ten for the 4-8 PM time period for 2021.

From figure 9, we once again observe that majority of the topics across Nine and Ten are sufficiently different due to orange being the predominant colour suggesting a JSD greater than 0.5. However, some notably similar topics are; 4 (Royals) against 3 (Royals) and 16 (Lockdown) against 7 (Lockdown).

3.4 Discussion of topics

As shown in the previous section, most of the topics derived were found to be different from each other. We can clearly deduce that this must be due to the channels broadcasting different programs across 4-8 PM time period, hence deriving distinct main topics. However, as previously mentioned, there is evidence in figure 1 and figure 2 of the underlying structure of the channel programming being very similar across all channels even with the different topics.

We want to now consider looking at the covid lockdown topic from each of the channels. As stated in the introduction, covid, and in particular covid lockdowns, has been a prevalent discussion point in the news throughout 2021. To help prevent the spread of the virus, as well as inform citizens of restrictions, we want to see similar and consistent coverage on covid across all channels throughout the year. To help explore this, consider the plot of the daily proportion of the covid lockdown topics for each of the channels in figure 10.

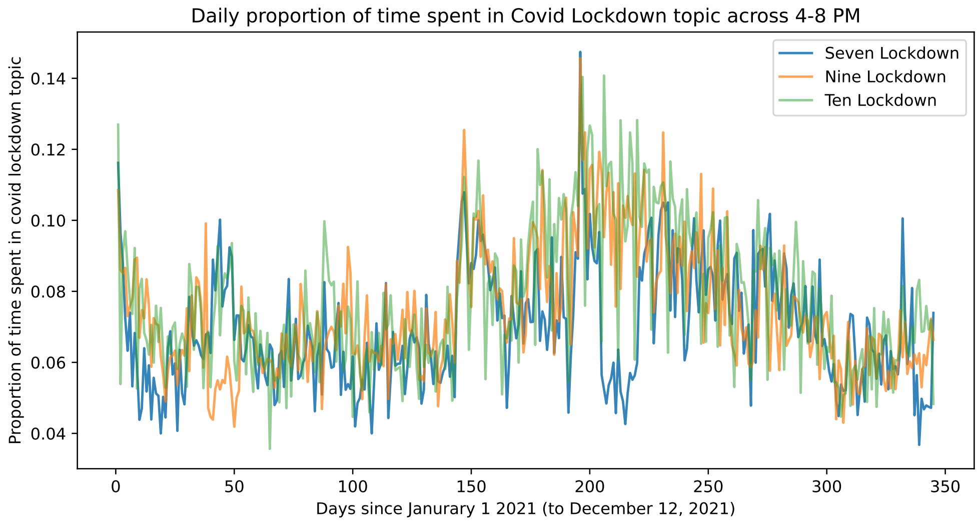

Figure 10: The daily proportion of the time spent in the covid lockdown topics across all the channels from January 1st 2021 to December 12th 2021.

From figure 10, we observe that the proportions of the covid lockdown topics across all the channels are similar, even exhibiting similar fluctuations throughout the yearly cycle. We also recall that the communication of the covid lockdown topics across the channels is consistent, as evident by the JSD in the previous section.

However, we do notice two drops in proportions for Nine, at 50 days (February), and Seven, at 200 days (late July), which are due to the Australian open and the Tokyo Olympics respectively, as discussed earlier. Interestingly, we can visibly see the point where media coverage focused more on covid lockdowns as Melbourne (June, 150 days), and Sydney (July, 180 days) started going into hard lockdowns. This is an important time for the media coverage of covid lockdowns to increase, as it is of critical importance that the general public oblige to the restrictions in place to prevent a longer lockdown.

4 Conclusion

In this report, we analysed TV media closed captions from the free to air commercial networks Seven, Nine and Ten, over a time period spanning from the 1st of January 2021, to the 12th of December 2021 in hourly increments. We first started by cleaning each of the data sets, which was an important procedure that helped to reduce the size of the data and also remove any redundancies present.

Next, we underwent some general analysis on each of the cleaned data sets. We found that both the total distrubution of words throughout the year across (figure 1), and the information entropy per hour (figure 2) across the channels were very similar. This may suggest an underlying programming structure that is shared across each of the channels.

Following on, we filtered the data set to only include key words from the 4-8 PM afternoon time period throughout the year and then did topic modelling on this new filtered data set. This revealed some interesting topics across all the channels, with examples being the Tokyo Olympics on Seven (table 2), the Tennis on Nine (table 3) and Master Chef on Ten (table 4). We also found that each of the channels had a covid lockdown topic, where the proportions throughout the year across all the channels was found to be very similar (figure 10).

Finally, we compared topics across each of the channels using a measure called the Jensen-Shannon Divergence. We found that most of the topics across each of the channels were sufficiently different from each other. However, the covid lockdown topics across each of the channels were shown to be quite similar (figures 7, 8, 9).

Although our work revealed interesting findings regarding the free to air commercial television channels, the scope of our research was limited due to time constraints for this project. There are many ways we can explore this data set in more detail, some of which include:

- Doing further topic modelling on a non-filtered data set with a larger amount of topics.
- Exploring the rise of covid and covid vaccination across the news since the start of the pandemic in early 2020, to the current day.
- Comparing how the news differs between a commercial network (seven) and a dedicated news network (ABC News 24).
- Doing sentiment analysis throughout the day to determine the happiest time period, and the saddest time period.

Acknowledgements

The author would once again like to acknowledge the contributions of fellow University of Adelaide student Ben Lang, for his help in storing all the data from the tveeder service for 2021 across all three channels, as well as the supervisor Professor Lewis Mitchell, for his guidance in this research project.

References

Blei, D 2012, 'Probabilistic Topic Models', *Communications of the ACM*, vol. 55, no. 4, pp. 77-88.

Lin, J 1991, 'Divergence measures based on the Shannon entropy', *IEEE TRANSACTIONS ON INFORMATION THEORY*, vol. 37, no. 1, pp. 147–149.

Kapadia, S 2019, 'Evaluate Topic Models: Latent Dirichlet Allocation (LDA)', *towards data science*, accessed at: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0.

Mach, K.J, Salas Reyes, R, Pentz, B et al 2021, 'News media coverage of COVID-19 public health and policy information', *Humanities and Social Sciences Communications*, vol. 8, no. 220.

Shannon, C 1948, 'A mathematical theory of communication', *The Bell System Technical Journal*, vol. 27, pp. 379–423.

Su, Z, McDonnell, D, Wen, J et al 2021, 'Mental health consequences of COVID-19 media coverage: the need for effective crisis communication practices', *Globalization and Health*, vol. 17, no. 4

Zhou, N 2021, 'Meet the man behind Tveeder, the no-frills live TV transcript that became an Australian media hero', *the Guardian*, accessed at: https://www.theguardian.com/media/2021/aug/21/meet-the-man-behind-tveeder-the-no-frills-live-tv-transcript-that-became-an-australian-media-hero

Appendix

Topic No.	Top 5 Common Words	Likely Topic
1	goal, ball, kick, player, ground	AFL
2	vaccine, health, royal, prince, queen	Royals
3	race, horse, group, track, winner	Horse Racing
4	case, lockdown, covid, site, health	Covid lockdown
5	bleep, mate, staff, alright, hand	**Unknown**
6	shower, rain, sunshine, afternoon, wind	Weather
7	garden, upbeat, beautiful, perfect, plant	Gardening
8	officer, police, road, passenger, driver	Road show
9	ball, wicket, player, cricket, shot	Cricket
10	gold, medal, olympic, tokyo, aussie	Tokyo Olympics
11	police, covid, government, state, country	Covid / Government
12	girl, voice, harry, lady, heart	**Unknown**
13	money, christmas, child, business, step	Christmas / Business
14	correct, question, answer, chase, chaser	The Chaser
15	brother, applause, challenge, housemate, laughter	Reality show
16	sigh, alright, talk, dean, fine	**Unknown**

Table 5: All 16 topics derived for the filtered Seven data set. The 'Likely Theme' column was determined by the author, with an **Unknown** given if it is not clear what the topic could represent.

Topic No.	Top 5 Common Words	Likely Theme
1	police, christmas, covid, vaccination, child	General covid
2	police, president, hotel, flight, quarantine	**Unknown**
3	test, drug, police, driver, alright	Road show
4	government, royal, vaccine, prince, queen	Royals
5	drop, zone, applause, final, jack	**Unknown**
6	question, applause, lock, eddie, correct	Millionaire Hotseat
7	relationship, chuckle, kind, miss, party	**Unknown**
8	girl, animal, child, parent, alright	**Unknown**
9	shower, degree, afternoon, sydney, wind	Weather
10	beautiful, garden, water, food, area	Gardening
11	room, block, space, bedroom, mark	The Block
12	match, court, player, serve, tennis	Tennis
13	business, affair, maximum, degree, price	Business
14	player, line, ball, season, final	Sport
15	laughter, build, alright, challenge, idea	**Unknown**
16	lockdown, vaccine, covid, health, site	Covid lockdown

Table 6: All 16 topics derived for the filtered Nine data set. The 'Likely Theme' column was determined by the author, with an **Unknown** given if it is not clear what the topic could represent.

Topic No.	Top 5 Common Words	Likely Theme
1	mate, water, beach, patient, alright	Bondi Rescue
2	liam, steffy, carter, eric, quinn	**Unknown**
3	minister, royal, price, queen, prime	Royals
4	star, laugh, mate, laughter, celebrity	Reality show
5	project, story, question, government, break	The Project
6	laugh, alright, race, brooke, date	**Unknown**
7	lockdown, sydney, state, heath, south	Covid lockdown
8	vaccine, government, quarantine, hotel, health	Vaccination
9	game, tomorrow, degree, shower, player	Sport / Weather
10	chris, beautiful, amanda, space, barry	The Living Room
11	south, sydney, police, coast, rain	Sydney
12	voiceover, race, song, horse, target	**Unknown**
13	dish, cook, flavour, beautiful, mate	Master Chef
14	laughter, game, christmas, applause, song	**Unknown**
15	laugh, girl, meet, chuckle, sort	**Unknown**
16	game, george, brain, challenge, andrew	**Unknown**

Table 7: All 16 topics derived for the filtered Ten data set. The 'Likely Theme' column was determined by the author, with an **Unknown** given if it is not clear what the topic could represent.