Insight Report

ABSTRACT

In the analysis process I have chosen the third question which is “What were the top Australian news topics over the last decade, and what can they say about the national conversation”.

The underlying significance of this question is isolated into two sections, in which one-section manages top news subjects in Australia, and the other part of the question is to comprehend the national discussion based around the identified topics in the course of the most recent decade. The general stakeholder could be a news organization since it manages points identified with Australia, the more intrigued and appropriate stakeholder is the Australian Press Council (APC). Furthermore, the insight gathered from the analysis is likely more suitable for the selected stakeholder to implement an effective strategy(Council).

The Australian Press Council is considered a key stakeholder for this analysis because the council is responsible for commencing healthy media practice, promoting access to information to the community through public interest, and allowing the freedom of expression through media By the constitution. Since the council deals with the policy matter within its areas of interest. This analysis will help understand the media behavior in Australia over the last decade by analyzing the topics.

PROBLEM SPACE

AUSTRALIAN PRESS COUNCIL

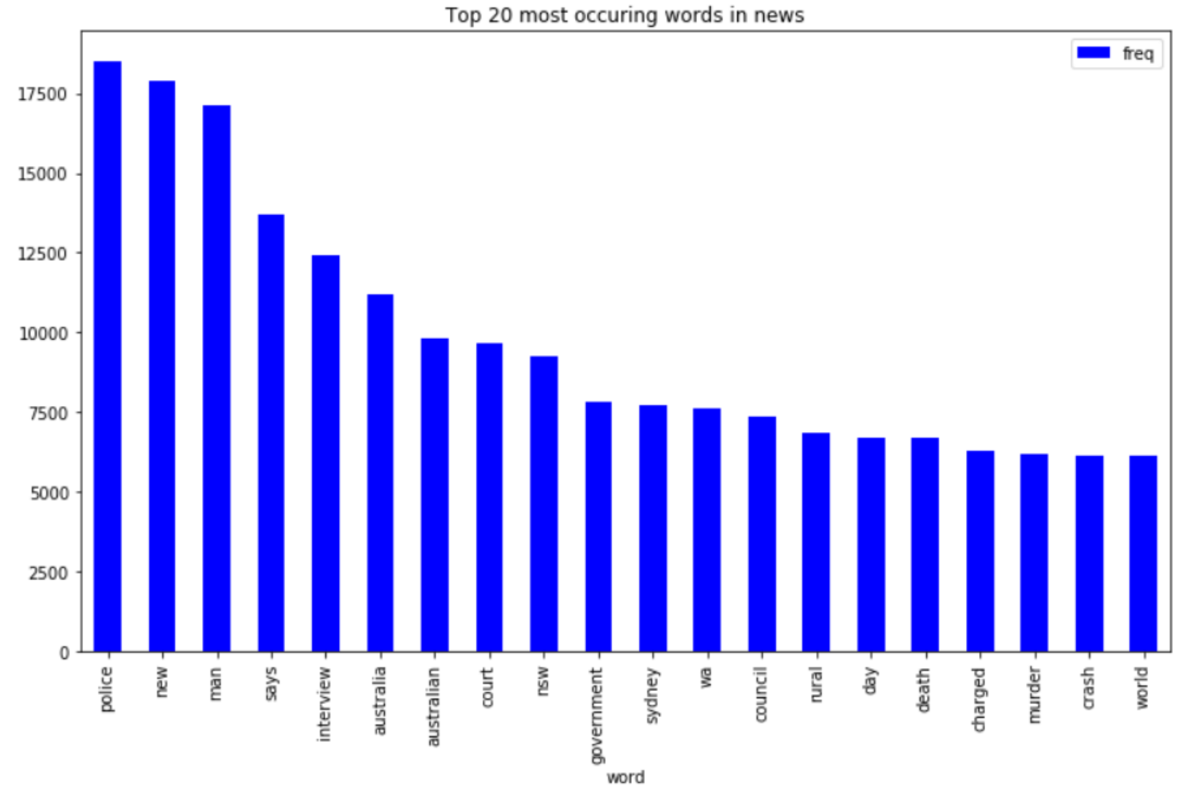
The term news agency is characterized as an association that gathers the news related data and circulates it to newspapers or telecasters. Nonetheless, now and again the circulation endeavors may be deceptive in the discoveries and give news that isn't moral in thought. Regardless of this, applicable and the significant news must be conveyed to its clients. Because of this type of deception Australian Press Council was framed. ACP deals in handling the complaints and concerns about the material in journals, newspapers, and magazines, ensuring the cause and concerns of the readers, promoting freedom of speech, etc(Council).

When it comes to the news reporting there have been several guidelines issued by the council to provide the news to viewers in the systematic format. Hence, this data analysis will help the Australian Press Council to analyze weather the issued code of ethics was followed by the news agency by examining the conversation. Because the ethics state that the publishers and editors are responsible for using text, headlines, photographs, and all the various forms of the published material. Also, this data analysis plays a vital role in understanding the strengths and effectiveness of the code of ethics. Also, providing the room for improvement in revising the existing code ethics or develop a new one.

ANALYSIS

The process of data analysis started from collecting the relative data from Kaggle. The dataset name is defined as ABC news. The Text analysis approach is implemented and the necessary libraries are imported. The data is loaded into the pandas data frame. The process of examining the data is undertaken, in which the data is examined for any null values present in the data set and then replacing those null values. There were no null values in this dataset. During data processing, I converted all the upper-case letters to lower case letters which were present in “headline_text” to avoid uneven analysis of the dataset. The word tokenize method is implemented from nltk to split the news data into a single set of words because it helps in better text understanding in machine learning applications. And is also helpful in the text cleaning process. Then the porter stemmer process is used because it reduces the words which are related to a common stem and also removes morphological words in English. The stemming is an important part of the analysis because it will help in answering the significance of this question. The words which are stemmed are stored in the new column called "stemmed_words". Removing of stop words is important because they add very little to the topic. Moving forward, I created the set of words called "stop_words" because creating a set is a much rapid process in removing the stopwords. To store the stemmed words after applying the stop word removal method, the separate column is created named "processed_stem". After the implementation of stemming and stopword removal, I have rejoined the words and stored them in the column named "processed_convo" which means processed conversation.

After loading and processing the data, the data analysis starts for selecting the year from 2010 to 2019 because the questions state the top Australian topics over the last decade. Hence, the data is filtered according to the mentioned date. The below diagram shows the top 20 most occurring news topics over the last decade. With police being the highest frequency.



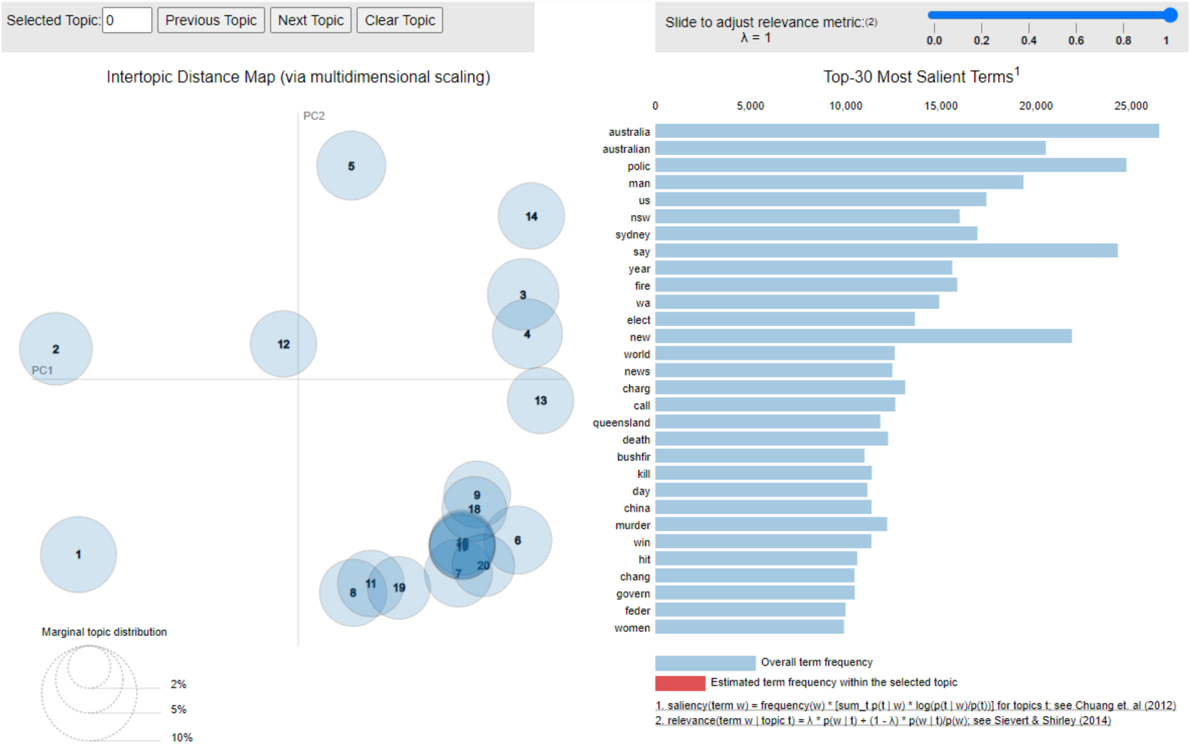
I created a vocabulary of words present in the "processed_convo" and assigned the token id for each word. This created a vocabulary that helped in creating the corpus. This corpus is given to TF-IDF. The TF-IDF algorithm is used because it allows numerical statistics which is important in understanding the words in documents collected in a corpus. It is mostly used as a weighting factor and tf-IDF value increases proportionally whenever the numerous words appear in a document and are offset by the number of documents in the corpus which consist of those sets of words.

The LDA model is used to discover the topics that are abstract and occur in the collection of documents known as topic modeling. Here, we can classify a text in a document to a particular topic. Therefore with the help of LDA, I was able to build a topic per document and word per topic model as distributions of LDA.

Below, we can see in topic 2 of our document the “Sydney” topic may consist of words “win” which has a 57% probability or “plan” which has a 35% probability. This means based on the given probabilities it will fill the document word slots.

```
Topic2
0.086*"sydney" + 0.057*"win" + 0.051*"abc" + 0.035*"plan" + 0.027*"talk
" + 0.026*"drug" + 0.026*"commun" + 0.025*"big" + 0.024*"polit" + 0.024
*"reveal" + 0.021*"white" + 0.019*"council" + 0.018*"give" + 0.016*"thi
" + 0.016*"green"
```

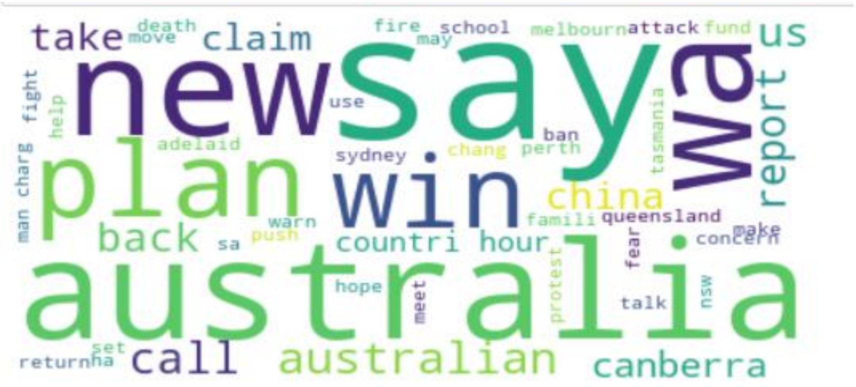
The visualization of pylDA is given below



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

pyLDAvis is implemented because it allows the users to interact with the topics in the presented topic model which has been fit to a corpus of text data. The global topics are displayed on the left and the term bar charts are displayed on the right. After linking the selections between the cluster data that gathered in the bottom right-hand side intertopic distance map, we can identify the top 30 topics. That cluster contains the topic such as “bushfire”, “Canberra”, “Australia” and “election”. Based on this cluster of data we can say that over the last decade the conversation is related to the above-mentioned topics.

Furthermore, with the help of the word cloud, I have displayed the top 50 set words present in the “processed_convo”.



INSIGHTS

AUSTRALIAN PRESS COUNCIL

The Australian press council conducts events and educational activities every year. The council assists university journalism lectures and various other forms of materials and instructions for case studies. These case studies involve various subjects matter related to the news. Hence, this data analysis can be used as one of the case studies for educational purposes. Also, ACP conducts numerous events. ACP can conduct one such event by inviting not only the mass media but also Australians to engage in a conversation about what Australians were interested in reading in the last decade. The topic modeling process provided with insight on understanding the most relevant topics and with the help of pyLDA visualization, we can see the most relevant topics over the last decade were “bushfire”, “Canberra”, “election” and “Australia”. Based on this, we can analyze the national conversation. Furthermore, ACP can also fact check some of the headlines related to the topic and check whether the news outlet was following the standard principles.

ETHICAL CONSIDERATION

The collected information from the data cannot determine the accuracy and whether it consists of factual information. This is essential because in the “ABC news dataset” only provides the headlines and sometime these headlines could be misleading and the actual content might give some different meaning. Thus, this analysis could not be taken into consideration to explain the detailed national conversation of Australia. Also, analysis does not disclose conflicts of interest that could affect the conversation. To understand this, we need a detailed data set with all the textual information to conduct analysis. The data set selected for analysis only provides information regarding the ABC news network and based on this data we cannot justify the national conversation of Australia. Thus, to achieve the more accurate results we need to look into more data from various news organisations in Australia.

The implementation of TF-IDF was enough for visualization. But implementing the LDA topic modeling algorithm helped in addressing the shortcomings of TF-IDF. The LDA uses probabilistic model which is useful in estimating the probability distributions for creating topics in words and documents. Here, the topic selection is based on its probability weightage and provides only words for the topics. And the problem may arise if the LDA is run several times because each time it runs the words probability could change. Thus, resulting in uneven word interpretation.

PRINCIPLES

To check the accuracy and facts of the data we need to examine the sources and check the references. To do this, we need to examine more data sets for various media outlets. For instance, the topics in ABC news could be used for comparison with other news channels. The data set used in the ABC news consists only of headings. Thus, collecting the whole article related to those headlines will improve in implementing better data analysis and will provide better data analysis results. The important aspect of finding the national conversation is to understand and examine the data for both positive and negative conversations or vice versa. For instance, democrats and republicans might not have same opinion but both might agree on some of the opinions. Thus, analysis of those opinions matters to find a gradual ground between the opposing parties. Therefore, this analysis is conducted in such a way that there is no consideration for being bias. Furthermore, extracting the articles of both the conversation and performing analysis could give us an in-depth critical analysis of national conversation. Although, the data set used here is solely based on news networks hence more such datasets from various other news networks could help more in understanding the Australian national conversation.

The advanced text analytics approach can be used for this data analysis process. The text categorization is on such a process that is powered by machine learning. Using text categorization allows the user to input the data and no manual implementation is necessary. The main advantage is that it will capture the importance of the word occurs in a text and provide better accuracy. Another approach is thematic analysis. In the thematic analysis, the themes are extracted from the text instead of simply categorizing the text. With the help of the thematic analysis, we can analyze the entire phrases which prove essential when dealing with data set such ABC news. Because through this process it is more reliable to analyze the conversation and could provide more elaborate information regarding the national conversation.

CONSEQUENCES

The Australian Press Council receives about 700 complaints every year(Council). Most of the complaints are related to correction or apology. The use of this method of data analysis will help the ACP in terms of reducing the complaints. For instance, ACP can now take look at this conversation over the decade and match the complaint ratio that was issued based on topics that were presented in this analysis. Doing this, the ACP will be able to understand on which topic they received the greatest number of complaints. And would also be able to conduct extensive research on why such topics were getting more complaints. Furthermore, this data analysis can help in either changing the regulations or creating more new regulations which would help the general public, journalist, and news media.

Coming towards the negative impact of this data analytics is that it solely focuses on single news organizations. And based on this organization alone the Australian Press council would not be able to decide to formulate new regulations or principles, because it may seem bias towards one media outlet. However, if the ACP takes any such decisions it may impact their credibility. If the ACP takes the decisions of censoring some of the words to reduce complaints, then the very core of ACP could be questioned which stands for freedom of expression.

Lastly, it is very important to take into consideration both the positive and negative aspects of this data analysis by the stakeholder and data analyst before making changes to regulations. In conclusion, more analysis should be done on various and different kinds of data sets.

REFERENCES

Council, A. P.: Australian Press Council.