Queensland University of Technology

Data Analytics for Strategic Decision Makers

IFN619

Insight Report

Chaitra Prathap (N10329510)

Table of Contents

Abstract

The purpose of the report is to analyse the insights retrieved from visualizating the datasets of ABCnews website. The report includes the technique used to retrieve the data and sort the top topics from the datasets. Critical analysis on the improvised visualisation is clearly described with graphical representation and the screenshots of the outcome in jupyter notebook. The problem space in the given question is analysed and give key insights to the stakeholder.

1.Background

This report provides a deeper understanding of the National conversation of top Australian news over the last decade (question 3 from 1B). Top Australia news mainly consists of police death, government calls, court charges covering all significant states likes QLD, NSW, VIC. This news dataset is visualized and categorized with top words. According to the requirements, I have reported the insights for the analysis and stated the significant context; which includes ethical considerations, principles and the potential consequences for people in different social contexts. From the feedback given, I have improvised the explanations and the visualization of abcnews by following LSA and LDA analysis technique. The chosen question for more in-depth analysis is stated below

1.1 Question

What were the top Australian news topics over the last decade, and what can these say about the national conversation?

2. Objective

Collecting published reports from the available primary news sources and categorising this topic using various analysing methods can develop a superior understanding of the context that the stakeholder covers in communicating this issue to the general public.

2.1 Methods

2.1.1 Data collection

The datasets collected are used in analytics and forecasted from big organisations, and government events with a volume of hundreds of articles reported. This data is retrieved from the following link (https://www.kaggle.com/therohk/million-headlines). The news dataset summarises a record of significant events from early 2003 to the end of 2019 with more focus on Australia. This includes the entire corpus of articles published by the abcnews website in the given time range. The article reports more than two hundred events per day, which covers international news and relatively all significant news are captured in this analysis. Searching through the keywords, you can see all the significant events that have influenced the last decade and how they have changed over time. For instance, the keywords include the war in Afghanistan, economic crisis, endless elections, climate crisis, terrorism, influential citizens, illegal activity and so on.

2.1.2 Data preparation

After considering the data sets, the data which was retrieved was CSV file where the date published was not structured. So, the data has to be organised in year-month-day format with separate columns in the excel sheet to proceed with the further cleaning in jupyter notebook.

2.1.3 Data cleaning

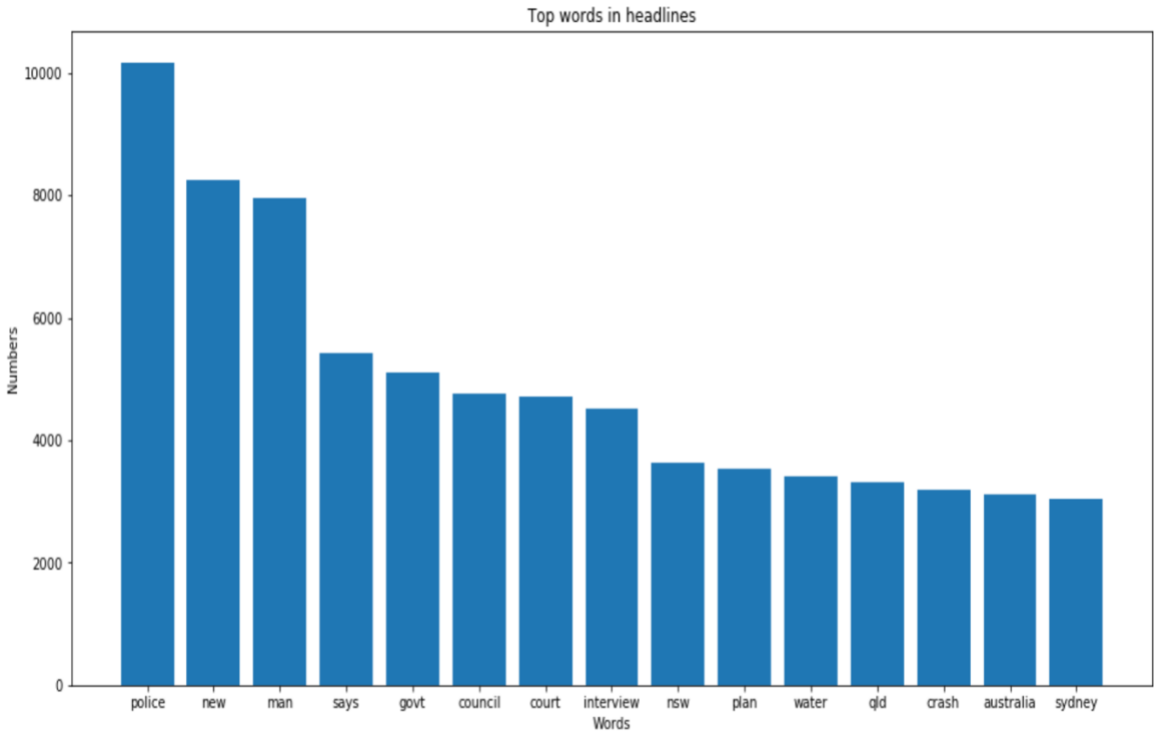
Here the large data as to go through a set of process to analysis. Firstly, to make sure data has no missing values; general extraction of data is performed. Secondly, as per the requirements, the data has to consider last decade values, so manipulation of data is performed. Further pre-processing steps are followed.

2.1.4 Critical analysis

For the above question, I have performed the following process in the jupyter notebook. To explore the data, data cleaning and pre-processing the abcnews datasets is performed; which gives the cleaned sets of column and rows of the data. As per the requirement, the data is manipulated or cut down to ten years by setting the period range parameter to the whole dataset. In order to plot the datasets, stop words were eliminated to avoid the errors in further analysis. Top words in this analysis are used to get very accurate counts of the words in the datasets.

Further pos_tags methods were also used to convert all the lists of each headline. In this analysis, the only pre-processing step required is the construction of features, under which we carry the dataset of text headlines and represent them in some solvable space. In practice, this means that each string is converted to a numerical vector. This can be done using SKLearn 's object Count Vectorizer; the datasets are sorted to top predicted topics of each headline, which retrieves the sorted top news topics. In the semantic analysis, the stop words remove the texts or words to get words which are useful for processing the data. Using Vader in sentimental analysis helps explicitly to pull out the list of sensitive data in the datasets and then the datasets are categorized by the type of words occurred (e.g., positive, negative and unbiased words). From the above method visualizing the datasets for top news makes it much easier, and the insights for all the visualization will be explained in detail. After which, using a bar chart, the relative magnitudes of all these categories can be easily visualized. t- distributed Stochastic Neighbour Embedding, which converts similarities between data points to probabilities and helps to minimize the high or low dimensional data. To implement clustering algorithm, I have used Latent Semantic Analysis and Latent Dirichlet Allocation to categorize the top headlines on the abcnews website.

2.2 Insights



From the above graph, the top words out of 15 cases from last decade in the headlines show police cases have been reported more than 10 thousand, and the court cases reports around 4500 cases and the least is the Sydney news which is around 3800 cases. The plan, water, Queensland, crashes and many other cases are at constant range of 3800 cases over last decade.

[318]:
```
sample = last_decade['HEADLINE_TEXT'][107880]
print(sample)
print('Sentiment: ')
print(sentiment_value(sample))
```

```
jetstar admits stranded passengers should have
Sentiment:
0.3
```

The above screenshots show the sentimental analysis of the 107880 data event, where the sentimental value for that event is 0.3 and this results as strong positive word in the event published.

```
]: sample1 = last_decade['HEADLINE_TEXT'][50394]
print(sample1)
print('Sentiment: ')
print(sentiment_value(sample1))
```

```
children suffering on hospital waiting lists
Sentiment:
-0.5
```

The above screenshots show the sentimental analysis of the 50394 data event, where the sentimental value for that event is -0.5 and this results as negative word in the event published.

```
: sample2 = last_decade['HEADLINE_TEXT'][100534]
print(sample2)
print('Sentiment: ')
print(sentiment_value(sample2))
```

```
grey nomads in tas not so grey
Sentiment:
0.1
```

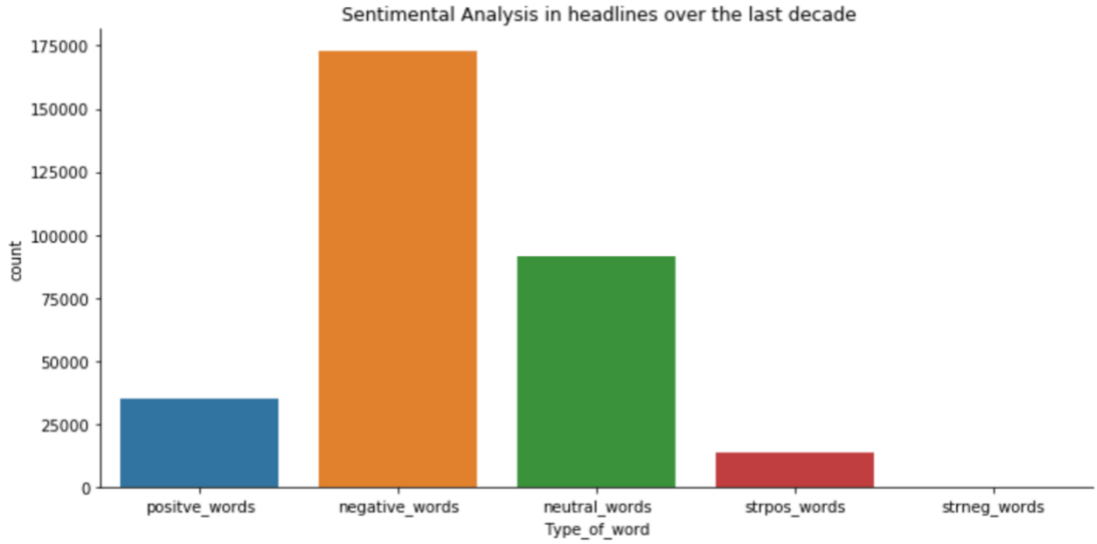
The above screenshots show the sentimental analysis of the 100534 data event, where the sentimental value for that event is 0.1 and this results as positive word in the event published.

```
[340]: sample = last_decade['HEADLINE_TEXT'][109880]
print(sample)
print('Sentiment: ')
print(sentiment_value(sample))
```

```
ramos horta addresses public
Sentiment:
0.0
```

The above screenshots show the sentimental analysis of the 109880 data event, where the sentimental value for that event is 0.0 and this results as neutral word in the event published.

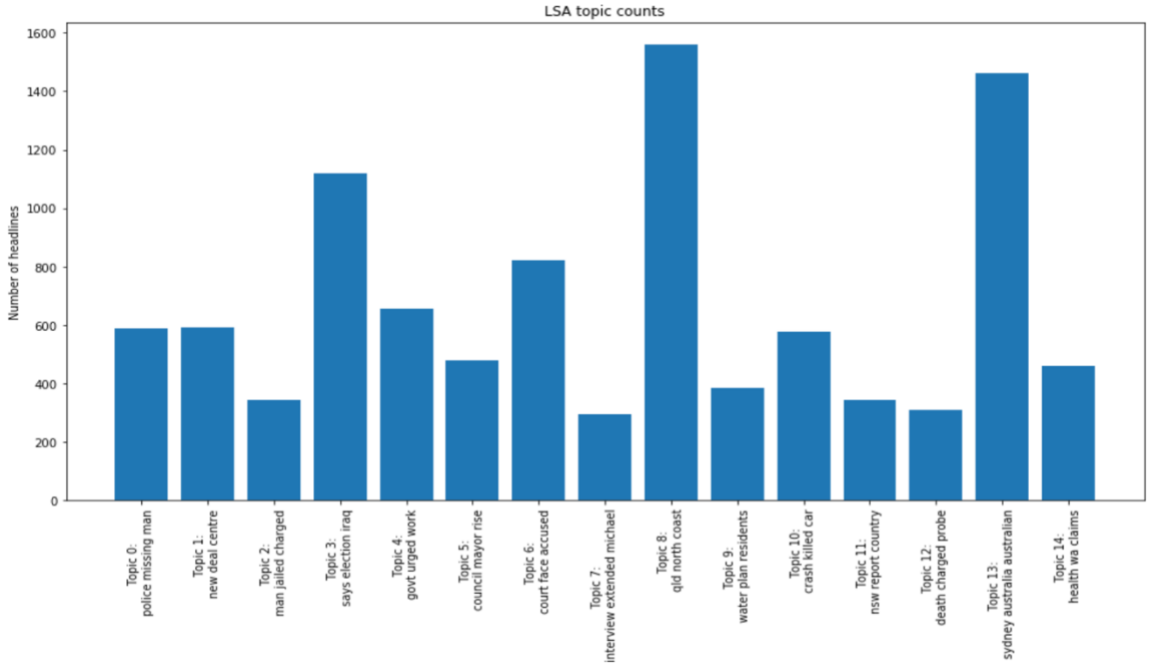
```
338]: <seaborn.axisgrid.FacetGrid at 0x7f2a270c85d0>
```



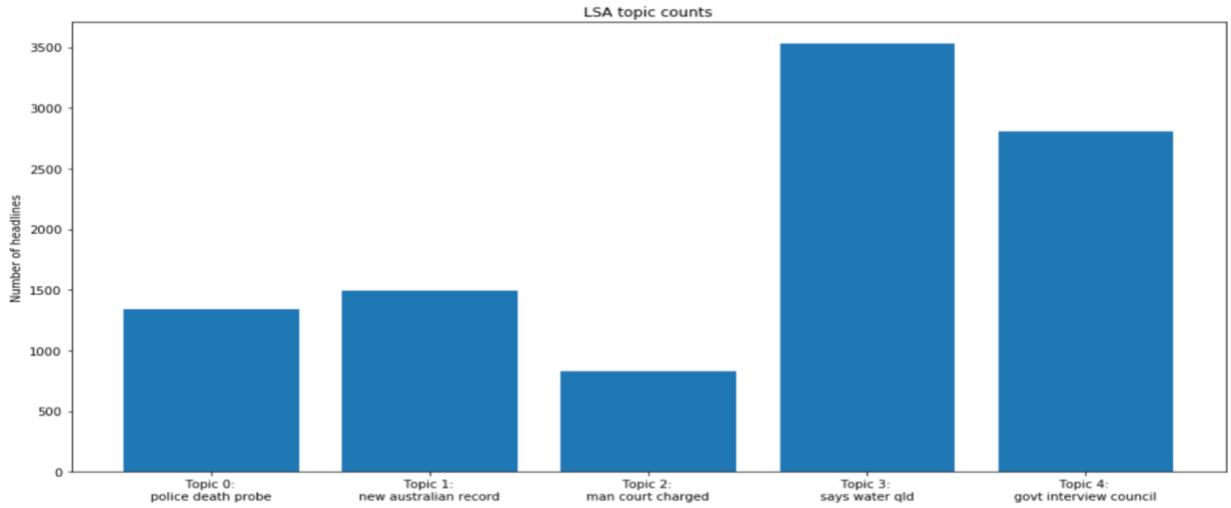
The above graph shows the type of words occurred in the datasets and the analysed outcome shows that there were no strong negative words in the news published over last decade. From the sentimental analysis, the positive news is less reported in the Australian news compared to negative and unbiased news.

```
Topic 1:  police missing probe search man hunt station investigate murder body
Topic 2:  new centre deal year president laws security workers party opens
Topic 3:  man charged jailed murder child guilty years court jail dies
Topic 4:  says minister funding iraq tax budget group changes opposition jobs
Topic 5:  govt urged work abc funds public urges defends boost vic
Topic 6:  council rise mayor rates gets power city change rate considers
Topic 7:  court accused face high case charges drug told trial title
Topic 8:  interview extended michael john nrl peter speaks tim matt josh
Topic 9:  qld health north coast rural election sa farmers national concerns
Topic 10:  water plan canberra murray act group protest dam seeks residents
Topic 11:  crash killed car woman dead dies school hospital driver melbourne
Topic 12:  death claims charged calls attack probe home inquiry baby toll
Topic 13:  nsw wa report country hour nrn 2014 13 friday drum
Topic 14:  east flooded korea artfair revellers tension extends laying poetic vegemite
Topic 15:  australia sydney australian world cup win final record day set
```

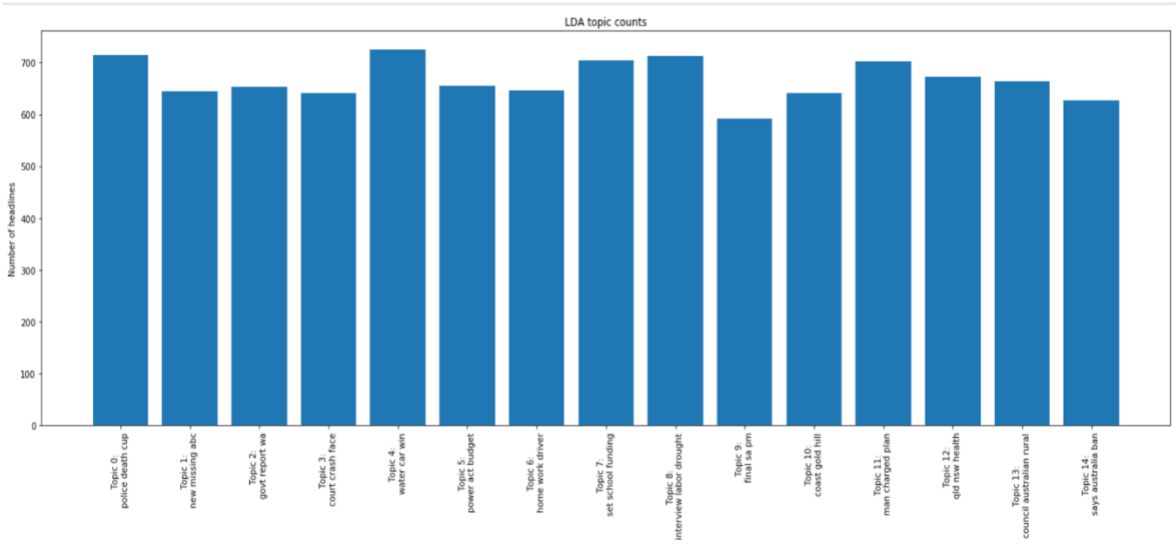
The above screenshot shows the top 15 topics retrieved from the datasets, where in the police missing cases tops the list and Australian world cup records is the top 15th topic sorted from the analyser.



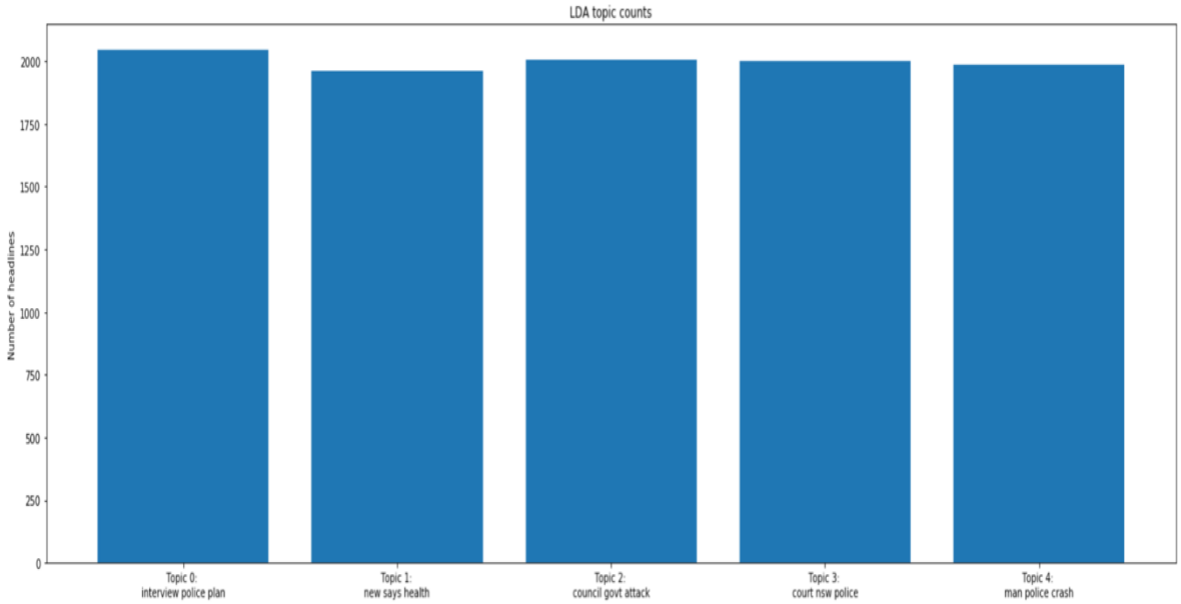
The graph shows top 15 topics of Australian news over decade and the graph is shortened to 5 topics for proper analysis. Using Latent Semantic Analysis the top events are shortlisted to this.



The above graph which is focused only on the top 5 topics of Australian news reported are police death Sydney, man court charged and interview minister with average counts, but new council plans and government NSW water cases stand out in the Visualisation. After pre-processing the data, the top topics in the data shows new council plans reached more than 4,000; police death in Sydney counts around 1,600, and the least is an interview with the minister, which is in top 5 of the Australian news. From the top topic's headlines, we can state that the most reported news includes government and council and the police case.



The graph shows top 15 topics of Australian news over decade and the graph is shortened to 5 topics for proper analysis. Using Latent Dirichlet Allocation the top 15 events are shortlisted to this.



The above graph which is focused only on the top 5 topics of Australian news reported are police interviews, health plans, council government attack with average counts, and police crash and NSW police stand out in the Visualisation. After pre-processing the data using Latent Dirichlet Allocation, the top topics in the data shows constant number of cases from all the events.

From the above to techniques LSA which gives the reduced matrix of words fails to categorize the topic but LDA applies appropriate algorithm to categorizes the topic.

3.Results

3.1 Critical analysis

Stakeholders

Stakeholders are journalists, government news sources, financial analysts, media and public affairs groups, and other platforms through which the events are reported.
The insight report focuses on:

Media and Public Affairs Groups: News agencies, reporters and shareholders trying to make a profit by reporting news in media. Their ideas will only be focused on increasing the websites rates from reporting more events and advertising them to the general public.

3.1.1 Problem Space

- News embassies interview patterns.
- Media groups are always in favour of the government.
- Media's strategy to maintain substantial exposure to its platform.
- To reach younger generation attention.
- News is published, focusing the Australian citizen's interests.
- Understand market presence and chances of advertising in Australia.
- Media considers the risk of constant stream crashing out.
- To cover events across all the region, state, cities and the bushes.
- Media groups ensure mutual survival and report local news to local audiences.

3.1.2 Problems in the Space

- News articles about government are always subject to change to keep sensitive matters out of the public domain.
- The news headlines trends topic might affect public and government officials.
- Reports news on a daily basis, be they positive, neutral or a negative.
- The news agency will function across a network of zonal, national, or division reporters investigating what is going on at their regions. Some news agencies compromise the news for mutual survival with the competitors.

3.2 Ethical Considerations

3.2.1 Sources of Data

- Not all views of Australian government statistics and parliamentary legislative powers have been captured, so these critics are censored and directed by the government, resulting in biased views of tradition and ethics.
- Governmental alternatives are not necessarily used as a data source, excluded due to government command.

3.2.2 Culture values to consider

- The datasets involved particular age group/ gender; the analysis relies on the cause and impact of the activity.
- Censorship limits access to other datasets, such as constitutional powers, so We do not know how well controversial issues will affect the government's future movement.

3.2.3 Questionable Data Collection Practices

- Data collection is government regulated and controlled; this creates an issue of well the information is published?
- How much effect does each of these data sources have on Australian government? Does the Australian public go against the popular media critiques?
- Algorithmic bias – semantic analysis is implemented based on the news reported, and the right factors are applied to sort the events?
- The technique of data collection – how much of the Australian press data is published out of the data collected for analysis and visualization?
- How are events identified and investigated – for example, the youth groups from each event are grouped based on their age.

3.2.4 Social Relevance (Data Analytics Issue)

- The news journalists, in particular, knows the interests of citizens changes as the generation changes.

3.3 Principles

3.3.1 Bias Identification

- Datasets must correctly represent the categories and responses from multiple viewpoints.
- Project an accurate view by publicising from both positive and negative improvements, including conflicts.
- We need to review, the questions asked to each group based on the event.
- We need to understand the journalist's dataset's context to understand what personal views they mainly portray.

3.3.2 Data Selection Strategy

- Journalists need to be aware of the Government's involvement with some of the data sources.
- Identify the sensitive groups suitably, categorise them into separate analysis and independently represent the groups to avoid false statements.
- Analysts also need to know how the Government handles and maintains sources of social media so that social media activities help to understand the semantic analysis.

3.4 Consequences

- If Bias Identification is not clarified, experts could manipulate a group that may not have witnessed the event.
- If the sensitive matters are out for the public, the phishers or rivalry group might misuse the information.

4. Conclusion

From the analysis above, this report describes the problem space in the questions by analysing each perspective of a stakeholder and how effective it is to consider the ethics and principles bend to the work.