Queensland University of Technology

Queensland University of Technology

ASSIGNMENT 2 -INSIGHT REPORT

DARSHEEL DESHPANDE
STUDENT NO :N10287213

Abstract:

News has become an important part of life in this modern world and has an impact on the life of people who are viewing it. This report includes a deep insight of the analysis carried out on data collected from the ABC news network. A critical analysis of the data analysis conducted on data has been done which includes ethical consideration, principals, and consequences of such kind of analysis. The report also includes a detailed description of stakeholders, the analysis conducted, and the technique used for analysis. An insight from the analysis is provided which would be beneficial for stakeholders. Ethical consideration section stats about the issues which can be related to data analysis. The principal section of the report gives solutions for an issue that is raised in the ethical consideration section. The analysis section describes the analysis which had been carried on data using tfidf and LDA. There are some points raised for consideration of stakeholder (news agency) for further using this analysis for changing their strategy.

Question

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

The significance of this question is that from this we would be able to find which were the news topic Australian prefer to hear and watch. we will get an insight into national conversion based on the analysis. The insight from the analysis is useful for the stakeholder that is a news agency to formulate an effective strategy for the future.

The stakeholder of this analysis is a news agency whose primary business is to present news to its audience and conduct debate based on issues and topics going around the world. The business of news agency depends on the viewership of channel, advertisement by producers, merchandise, etc. The insight from this analysis will help them to know which are the hot topic which viewers prefer to watch in debate and news. This will help them to increase their viewership and also subsequently help them to attract more advertisements from producers and merchandise.

STAKEHOLDERS PERSPECTIVE:

As an administrator of a news agency, it is necessary to keep the viewers engaged with the channel so that we can generate an income. The business of the news agency particularly depends on the viewership from the show, bulletins, and debate conducted on the topics which are going around the world. Depending on the viewership of channel investors, producers, and merchandise companies are attracted to it to advertise their products on our channels which have also become one of the major sources of generating business in this industry. With globalization, there are many competitors in the industry which makes the race for the viewership a hard task. So it is the most important thing in news agency business to keep our viewers intact, this can be done by showing and discussing topics that make value to them or in which they are prominently interested. Generally, the news which is presented in debate and bulletin are those which come on a day to day basis on events that took place locally or in the world which may have some or other impact. The selection of news to be presented is mostly dependant on the editor choice and the viewers are mostly not taken into consideration while showing certain news this can affect the viewership of the channel. The insight from the analysis especially focuses on finding the hot topics which the Australian audience is interested in. This can give the agency as well as the editor an idea regarding the expectations of viewers from their channel. This gives us an upper hand over the competitors in projecting such kinds of topics in our shows and telecast which may lead to an increase in our viewership as well as would help our finance.

NEWSAGENCY: PROBLEM SPACE

News agency shows news from various parts of the world and conducts debate related to issues taking place, gives information, and spread awareness regarding a certain situation. The business model or getting funding of the news agency is prominently dependent on the viewership of shows and debate and the advertisement which are shown during shows which are also indirectly dependant on the viewership of the channel. The news agency is regulated by the policies and guidelines set up by the government of the country. These guidelines include displaying of particular news as well as receiving funds. With the increase in the competitor, it has become very essential for the agency to maintain its base viewership and plan according to the expansion of the business.

Data analysis plays an important role in this field from getting information from sources and displaying statistics as a part of the news. Since this industry is mostly dependant on the viewership data analytics plays an important role in maintaining and expanding the viewership in this field based on the graphs and projections available. This can help the agency to make changes in strategy for building a good base in this industry and can also be beneficial to maintain the financial aspect of running the agency. Data analytics can also help to change the decision-making model in news agency generally decision on presenting news is taken by editors and journalist in the editorial team and viewers view are not consider but data analytics can be used for taking viewers views in consideration.

ANALYSIS:

The analysis was conducted using the TFidf and topic modelling algorithm of LDA. In the previous assignment, I had used tfidf on the same data I improved the analysis by using LDA. The analysis was conducted on data that was collected from Kaggle on the ABC news network. The data was from the year 2003 to 2019. First data was cleaned and pre-processed since we required data from 2010-2019 to get news headline for the decade data to data frame index of the dataframe was changed to date and data from the year 2010 to 2019 was selected and was given to another dataframe since we wanted the analytics limited to this decade only. First, the sentence in the headline_text was tokenized and then stopwords were removed from the sentence. Stopwords are the words like for, is, the, etc which are used in sentences but do not have much importance in this analysis. Stopword removal was used since most of the sentences have these words in it if we would have considered these words during our analysis then we would have got some stopwords are words with the highest frequency which would be not useful for our analysis. The output of stopwords was given to bigrams .bigrams can be said like two words that occur frequently and have meaning together words like hong_kong,new_york, These words which have meaning together would have appeared as two separate words if bigrams would have not been used which would have given a different result .output of bigram was given to lemmatization, lemmatization is a process which takes words with a different part of speech and normalized them. Words with the same meaning but having different parts of speech would have been counted as a different word which would complicated analysis and would have given different results to avoid that lemmatization was used. The output of lemmatization was given to a dictionary to form a vocabulary based on words in which each word will have a unique token id. This vocabulary was used to build a corpus. This corpus was given as an input to tfidf model. The tfidf model uses a corpus and vocabulary to determine the occurrence of a term in the document. The term is sorted based on the term frequency in the document. To visualize the tfidf result word cloud was used the input to the column which has been sorted depending on tfidf. The visualization displays the top 500 words based on the number of occurrences in the document as per tfidf model.

[22]:

To find the topics a topic modelling technique of LDA was selected the input to the LDA model was given as the vocabulary, the corpus which was created, and the number of topics that we want to display .20 topics were considered to be displayed. LDA considers keywords to form a particular topic and it used probability to stat the value of keyword for a particular topic. LDA never names topic. The topic can be named by human interpretation based on the probability value of words for a particular topic. In this case, I have interpreted some of the topics based on the value of the probability.

Human interpretation of the topic was done as follows:

This can be interpreted as:

(5,

'0.104*"election" + 0.083*"news" + 0.070*"abc" + 0.059*"canberra" + '

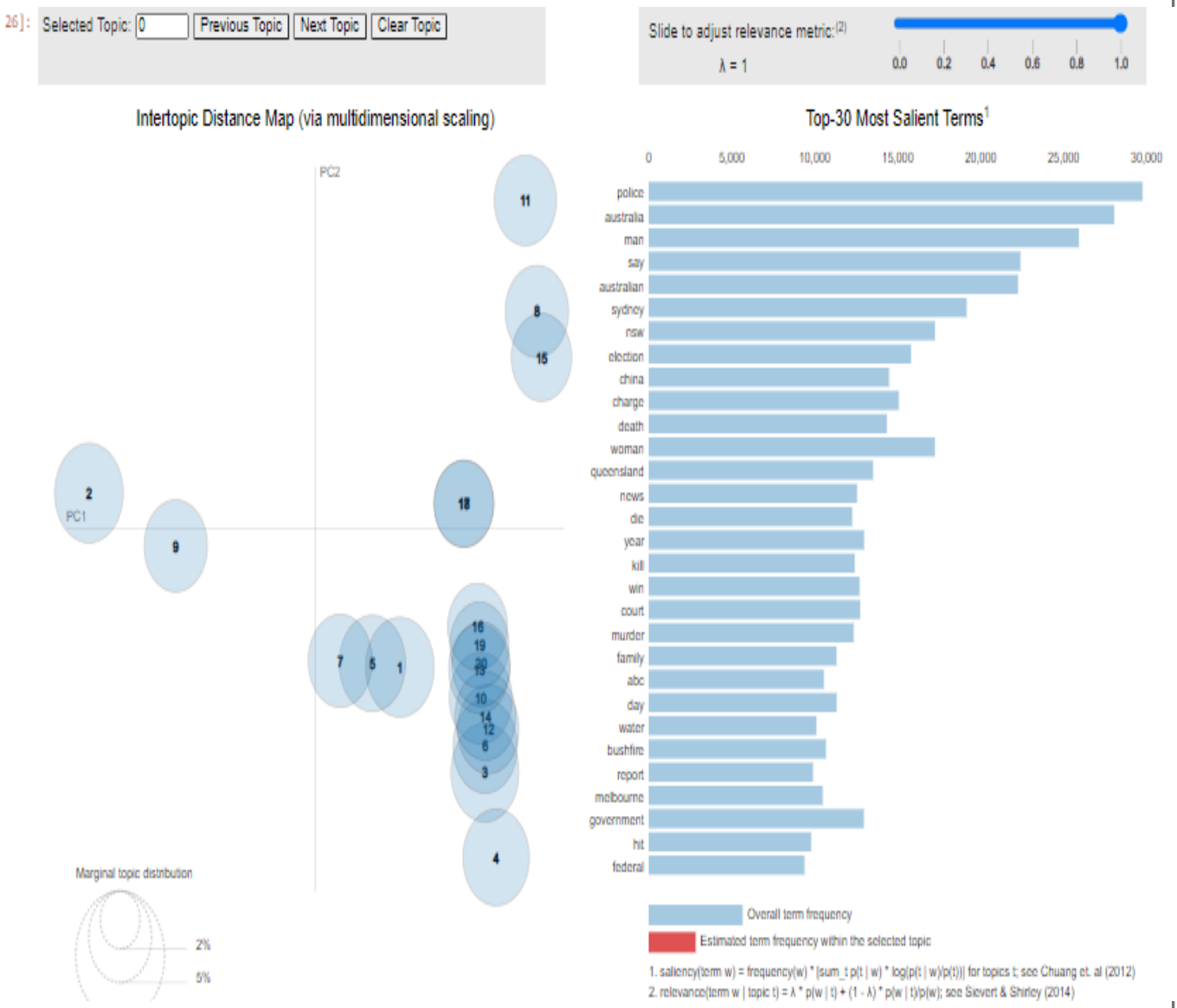
'0.042*"take" + 0.038*"warn" + 0.027*"city" + 0.027*"state" + '

'0.027*"emergency" + 0.024*"launch"')

it means that topic 5 is represented by following keywords and their probability related to the topic. The probability represents how the word is important to topic depending on the words you can guess the topic could be and summarise. for topic 5 we can summarise that on poll or politics. we take it as politics.

Likewise, the order topic was summarised.

To get more ideas regarding the topic pylDA visualization was used which displayed topics.



The topic is represented in a form of bubble in the left side plot. A model that has a greater number of a topic will have a cluster of a bubble at one side of the graph which can be seen from graph present. The topic which is there in these clusters is related to politics, governance, and policymaking and rules or law. The cluster consists of a majority of the top 20 topics so we can say that the majority of topics were related to politics and the national conversion of Australian in the last decade was based on political and governance issues.

INSIGHTS:

From the analysis, we can find 500 words that occur most in headlines from the year 2010 to 2019. They are being displayed in the word cloud using the TF-IDF algorithm. This word cloud gives an insight of popular words in the headline used by ABC news network.to get to know topic LDA was used we got the top 20 topics from the LDA which uses keywords to form topics. The topic was humanly interpreted based on the probability of keyword in a particular topic given by LDA. The interpretation of the topic is explained in the analysis section of this document. Based on interpretation we found that most of the topics are from topic type politics and governance. Other types of topics we found are related to type crime, trade, world news, etc. When we visualize the LDA output we can see that many topics form a cluster in a side of the left-hand side of graphs of visualization. From clicking on

the topic from clusters we got to know that all those topics are correlated and are from a similar topic type that is politics. Thus, from this cluster, we can say that the conversion among Australian is based on politics and governance topics.

INSIGHT FOR NEWS AGENCY:

According to the analysis, we can have an insight that the viewers of ABC news share interest in the topic related to political updates, government rules and regulations, and also news related to policy change. such kinds of analysis can help the news agency know what the targeted audience is interested in. From the analysis and visualization in both tfidf and LDA, it is clear that the Australian audience is more interested in local and national issues rather than an international issue. This insight can help news agencies in Australia to focus their strategy more on a national and local issue, however national issues like change in policy, rules, etc should get paid more attention and be the centre of communication. With globalization and growing competitors in the industry, this insight from the analysis will give the news agency upper hand over the others. Such insight can help news agency increase their viewership and attract more business and finance for their company.

Ethical consideration:

There are many ethical issues can be taken into consideration they are as follows:

1. Data collected:

The data which is collected from Kaggle is from only a single news network that may only show the interest of the viewers who view ABC news network. The Australian who view other network or who do not watch the news on television is not taken under consideration for this analysis. Data collected shows that the viewers view the headline but it does not speak about the positive or negative response for the headlines like if they wanted to watch that headline or they were just waiting for a different show or casually viewing what is going around in news. The data can be mostly from a particular region as a big city and developed region. The regional area which does not have accessibility is not taken under consideration. if such problem is present in the data the insight of the analysis will be affected. Since it will give wrong information regarding the topic which we are selected. Thus, implementing changes based on the insight of such kind of data can lead to wrong decisions and can affect the desired goal of a news agency.

2. Analysis issue:

The analysis of tfidf can be trusted for showing important words in headlines as the algorithm of tfidf is based on only consideration of term frequency and inverse term frequency in documents. The topic modelling algorithm of latent Dirichlet allocation does not give the name of the topic it only gives the topic and the keywords associated with it with the probability weightage of a particular word for that topic. The topic is being interpreted by human interpretation. The human interpretation may vary from person to person therefore the name given to the topic cannot be considered as reliable. Changes made to decision making the process of stakeholder-based on a wrong analysis can have an impact on the business of stakeholders. Also, investment in such kind of analysis would be wasted.

Principles:

The data collection policy should have been changed. Data from other channels and networks should be collected and the analysis for all the networks should be examined to get to know more about the preference of users. more variables should be added to the dataset which can specify if users choose for preferred show or liking or disliking of a particular type of shows. This would give an insight into how their positive and negative feedback for particular types of headlines or shows displayed on news. Data from other countries like new Zealand who have more contextual similarity can also be considered and compare data from Australia.

For the data analytics process, we can also use a technique like nmf (non-negative matrix factorization) which is a linear algebraic model that factorizes high dimensions vector to low dimension vectors. The output of this will be similar To LDA and sometimes better depending on the dataset. if the types of topics in the dataset are already known we can use the Artificial intelligence explainability process with the help of naïve Bayes text classification can consider which takes the words and sentences and tell us which words and sentences are from a particular topic from the type of topics given as input. This all is done with the help of the naïve Bayes algorithm which trains and test the data for text classifications.

Consequences:

consequences of selecting such kind of data are if the data does not go well for all the viewers and projection based on the data does not match the expected level then it can affect the decision-making strategy of the news agency. They may have a loss of resources and funding due to wrong strategies adapted based on the analysis of this data. If the algorithm or the human interpretation is done for getting the topic name proves to be misinterpreted It may also affect the decision-making process of a news agency. It may have a financial loss to the stakeholder for investment in such analysis and also loss of time and resources. There can be a loss of viewership for the news agency and which could affect their funding's from other sources and promotions. There could be changes made in the decision-making panel of the news agency and also some individuals can lose their jobs due to the low business of that particular agency. There would be damage to the reputation of the news agency.

This kind of insights on analysis and critical analysis can prove useful also to the printing press and newspaper industry which slowly are shifting to a web-based version. The insights can help to article new readers and youth on a social media platform or other platforms that are used by today's generation to increase their popularity and create a new base of readers.