

Data Analytics Project

This report will focus on contextualizing business data solution using R to solve the problems and improve business decision.

Overall, it looks for gaining insights from the text data.

Objectives

The objective of this report is to analyzing articles are as follows:

- Content summarization: there are lots of articles in the dataset, and this report will extract the most representative topics.
- Information retrieval: it can be used to searching for document based on topics which will help on search engines and content recommendation system.
- Document organization: this report will illustrate the organized articles with cluster and categories based.
- Content understanding: This analysis will help to understanding the subject matter into targeted analysis.

Overall, the major objective is to extracting meaningful insights from the article textual data.

Tools and technologies

R program language has been using of open-source environment for data analysis. Specially, TM (text mining) and (topicmodels) packages are used for preprocessing text data. And for visualization tidyverse has been used. Whereas, dplyr, tidyr and ggplot2. The reason to choose R is, it has lots of advantages for topic modeling analysis and it can fit into our goals of article analysis.

Data description

This dataset is taken from (ASADMAHMOOD, 2017) Kaggle. This data has news articles related to business and sports. It has included the place of article published. It has 4 column which are Articles, Data, Heading and News Type.

Article: Text having the news article and the place where it was published from

Heading: Text containing the heading of the news article.

Date: Date when the article was published.

News Type: Type of Article examples: business or sports

Data Preprocessing

Data cleaning

Firstly, I loaded the original data, which was 2692obs. 4 variable and now after handled the unique_data its 2585obs of 4 variables. The data was cleaning by following codes:

```
# Check for missing values and remove rows with missing data
# This ensures that rows with missing values are removed, as missing data
can lead to analysis issues.
clean_data <- na.omit(Articles)
# Handle duplicates
# Removing duplicate rows ensures each unique data point is counted only
once.
unique_data <- unique(clean_data)
```

After loading the data, get a quick overview of the dataset using these functions.

```
# Display the first few rows of the dataset
head(data)

# Summary statistics for numeric columns
summary(data)

# Structure of the dataset (variable names and data types)
```

```
str(data)
```

```
# Dimensions of the dataset (number of rows and columns)
```

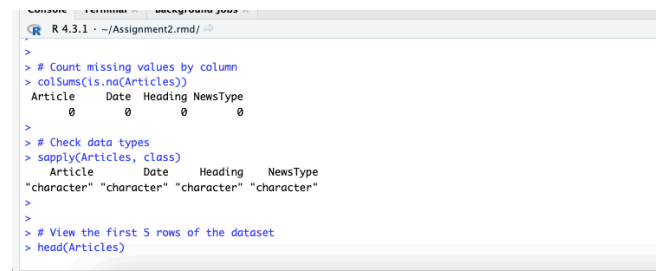
```
dim(data)
```

on the basis of above function, we should find **Missing Values:** Identify and count missing values in the dataset.

```
# Count missing values by column
```

```
colSums(is.na(data))
```

output was:



```
R 4.3.1 - ~/Assignment2.rmd /
>
> # Count missing values by column
> colSums(is.na(Articles))
Article      Date      Heading NewsType
      0         0         0         0
>
> # Check data types
> sapply(Articles, class)
Article      Date      Heading NewsType
"character" "character" "character" "character"
>
> # View the first 5 rows of the dataset
> head(Articles)
```

This is good news, as having missing data can sometimes complicate on data analysis.

Limitation

The limitation of extracting and analyzing patterns of data set from the news article are as follows:

- Overlapping topics: the data of news articles may cover multiple topics simultaneously and the result can be difficult to focus on single or clear topics.
- Influence of key words: as topic modeling focuses only in presence of keywords, the words may overrepresented and result can be biased.
- Evaluation challenges: while doing evaluation subjectivity in evaluation metric can be a limitation.
- Sentiment understanding: as topic model difficult to understand the semantic meaning of words the content or topic developed may lack semantic coherence.

Data visualization

LDA has been used for generative probabilistic model which are as follows:

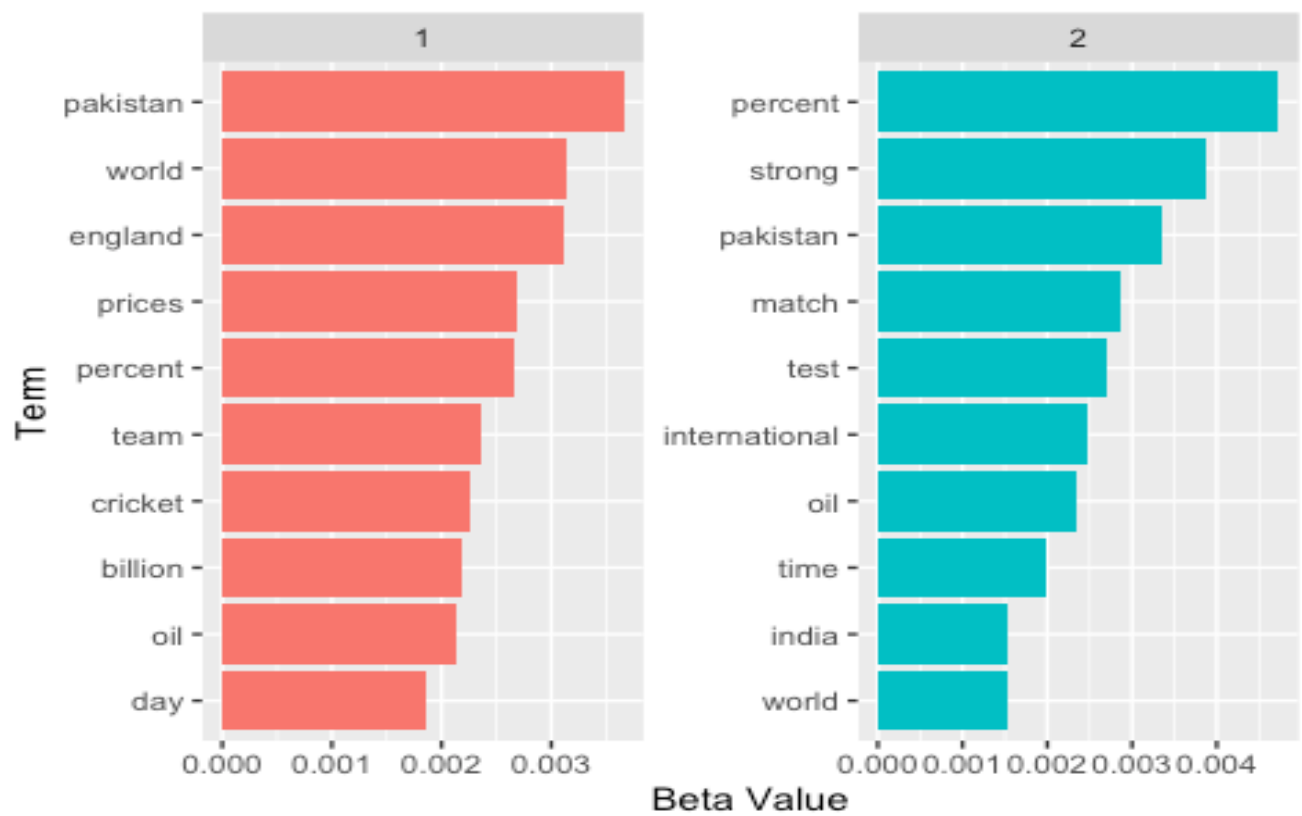
```
# Create a Document-Term Matrix (DTM) from your dataset
cleaned_articles_dtm <- cleaned_articles %>%
  unnest_tokens(word, Article) %>%
  count(Date, word) %>%
  cast_dtm(Date, word, n)

# Set a seed for reproducibility
set.seed(1234)

# Create an LDA topic model with 2 topics
lda_model <- LDA(cleaned_articles_dtm, k = 2)
```

This is used for the LDA function K=2 to create two topic model which will help to containing the full details of the model fit. This shows how topics and word are associated with articles.

Here, dplyrs slice_max has been used to find the most common each topic. The chart shown as below:



The above visualization represents the terms that are most common within each topic. Whereas, two topics are extracted from the articles. The most common word in topic 1 are Pakistan, world, England, prices, percent and so on which shows that it may illustrates the sports and business news. On the other hand, most common word in topic 2 are percent, strong, Pakistan, match, test, international and so on which may present the sports and financial news. One word which is noticeable is Pakistan in each topic which has common theme and it can reflect its diverse presence in the dataset.

As an alternative, this data set also consider the greatest difference in between topic 1 and 2.

```
> beta_wide
# A tibble: 55 × 4
  term      topic1  topic2 log_ratio
<chr>      <dbl>   <dbl>   <dbl>
```

```

1 country    0.0000718 0.00177      4.62
2 government 0.000454  0.00159      1.81
3 karachi    0.0000674 0.00102      3.91
4 prices     0.00216  0.000968    -1.16
5 1          0.000711 0.00111      0.648
6 2015       0.000101 0.00107      3.40
7 2016       0.000299 0.00130      2.12
8 africa     0.00101  0.000105    -3.25
9 asia       0.00102  0.000115    -3.15
10 australia 0.000441  0.00110      1.32
# ⓘ 45 more rows

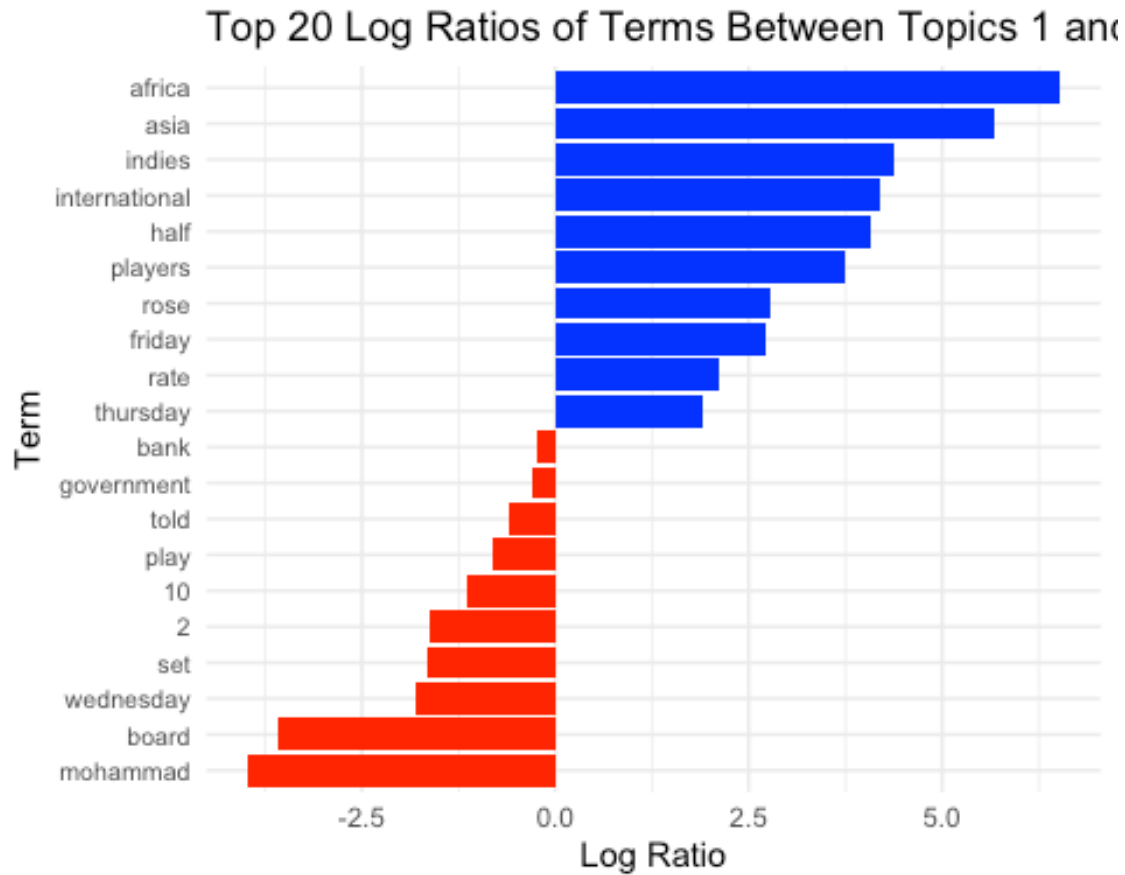
```

```
# ⓘ Use `print(n = ...)` to see more rows
```

The above table represent the associated with two topics in LDA topic model. The topic 1 and 2 shows the probability of each term being associated each other. The higher the value stronger the association with particular topic. The “Country” has a very low probability for topic 1 with 0.0000718 and higher probability for topic2 with 0.00177 which proves that it is more closely related to topic2.

In the row for "prices," the log ratio is negative (-1.16), indicating that this term is more strongly associated with topic2. Conversely, terms with positive log ratios are more strongly associated with topic1 and so on.

The following figure shows the words with the greatest differences between the two topics.



In above figures, it can be seen that the word most common in topic 2 are Africa, Asia, indies, international and so on which are related to sports. In topic 2, Mahammad, board, Wednesday, set are most common characterized by their name and business. Overall, it proves that the algorithm identified related to business and sports news.

Model Evaluation

The following coding represent the model evaluation with perplexity score:

```
#model evaluation  
# Calculate the log-likelihood of the model on the data frame 'cleaned_articles_dtm'  
log_likelihood <- logLik(lda_model, cleaned_articles_dtm)  
# Calculate the perplexity  
perplexity <- exp (-log_likelihood / sum(cleaned_articles_dtm))  
# Print the perplexity score  
cat ("Perplexity:", perplexity, "\n")  
Perplexity: 1655.059
```

The output of perplexity score got 1655.059. It illustrates a measure of how well this data pretrained LDA topic model is able to predict the words of dataset. As an interpretation of perplexity, lower values are better which indicates that the LDA model is better at predicting the words in dataset.

In addition, this perplexity score provides valuable insights into the model's ability to generalize from its training data to new, previously unseen text. While this score is a valuable component of our model evaluation, it should be considered holistically with other relevant factors to make informed decisions about the model's alignment with our specific project objectives.

Conclusion

To sum up the report, this project has provided the in-depth understanding of business decision making using real-word business data. The output of topic modeling and process of understanding of the dataset through rigorous data wrangling, visualization and exploratory analysis, it has help to understand the content of articles and that will help to make good decision making.

The visualization of data has given the communication of finding effectively with complex information in a clear and accessible manner with whether its sport or business-related topic. The communication and presentation of insights derived from the analysis is a vital aspect of this project.

References

Anderson, C., 2015. *Creating a data-driven organization: Practical advice from the trenches*. " O'Reilly Media, Inc."

ASADMAHMOOD, 2017. *Kaggle*. [Online]

Available at: <https://www.kaggle.com/datasets/asad1m9a9h6mood/news-articles>

[Accessed 1 september 2023].

Barde, B.V. and Bainwad, A.M., 2017, June. An overview of topic modeling methods and tools. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 745-750). IEEE.

Silge, J. and Robinson, D., 2017. *Text mining with R: A tidy approach*. " O'Reilly Media, Inc."

Vayansky, I. and Kumar, S.A., 2020. A review of topic modeling methods. *Information Systems*, 94, p.101582.