REPORT ASSIGNMENT 2:-

-By Shirish Shekhar Jha
20257
DSE

TITLE

EXPLORATORY DATA ANALYSIS OF TEXT DATA: ANALYSING NEWS DATA TO IDENTIFY TOPICS THAT WERE MORE IN DISCUSSION

OVERVIEW:

We have been looking to analyze the text data of the news headlines to identify the topics in the heat of discussion. The variables sought for are major words being used to analyze the issues.

RESEARCH QUESTION;

Sentimental analysis is a hot topic for now. Since people often present their views either in the form of text or in audio format, it is much more convenient to know about people's opinions using text data. Using text data analytics, we can infer what are the current topics to be aware of and how the current topics are impacting the daily lives of people.

References

https://towardsdatascience.com/nlp-part-3-exploratory-data-analysis-of-text-data-1caa8ab3f79d
https://towardsdatascience.com/preprocessing-text-data-using-python-576206753c28
https://www.analyticsvidhya.com/blog/2020/04/beginners-guide-exploratory-data-analysis-text-data/
https://huggingface.co/tasks/text-classification
https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/ch01.html

Hypothesis:

The appearance of words like Not, angry, War, against, etc. tells us that the headlines are negative and by seeing the frequency of some of the words we can make out the topic about which it is being discussed about.

Dataset:

Dataset Name:

A Million News Headlines

Dataset Link:

https://www.kaggle.com/datasets/therohk/million-headlines

Dataset Description:

This contains data of news headlines published over a period of nineteen years. Sourced from the reputable Australian news source ABC (Australian Broadcasting Corporation)
I look at this news dataset as a summarised historical record of noteworthy events in the globe from early-2003 to end-2021 with a more granular focus on Australia.
This includes the entire corpus of articles published by the ABC website in the given time range. With a volume of two hundred articles each day and a good focus on international news, we can be fairly certain that every event of significance has been captured here.
Digging into the keywords, one can see all the important episodes shaping the last decade and how they evolved over time. For example: financial crisis, iraq war, multiple elections, ecological disasters, terrorism, famous people, local crimes et

Number of variables: