Index No 1: EG/2020/4076

Index No 2: EG/2020/4077

Index No 3: EG/2020/4342

Topic – “Sentiment and Engagement Analysis on New York Times Comments”

Data set - New York Times Comments

Source code: AhamedMinhaj456/Sentiment-and-Engagement-Analysis-on-New-York-Times-Comments-with-Hadoop-MapReduce

Summary of the Results:

The MapReduce sentiment analysis was successfully executed on the *New York Times Comments* dataset using a custom-built Hadoop job. A sentiment lexicon was used to classify words in each comment as either positive or negative**.** This result clearly demonstrates that the comments examined from the dataset tended to be positive. The substantial difference indicates that, on the whole, users were expressing positive thoughts or responses in their comments.

The significant discrepancy between favorable and unfavorable opinions might be a reflection of the article or articles' content or the prevailing discourse at the time.

Although successful, the sentiment lexicon-based method might miss neutral or context-dependent sentiment, which could be enhanced in subsequent research by applying machine learning or natural language processing techniques.

Observations & Suggestions:

The MapReduce model proved effective for processing a sizable real-world dataset which is New York Times user comments on a single-node Hadoop setup. It illustrated Hadoop's ability to handle text-based sentiment classification tasks through scalable, parallel processing. The performance was sufficient for count-based aggregation, and the analysis successfully revealed that the majority of user comments carried a positive sentiment**.**

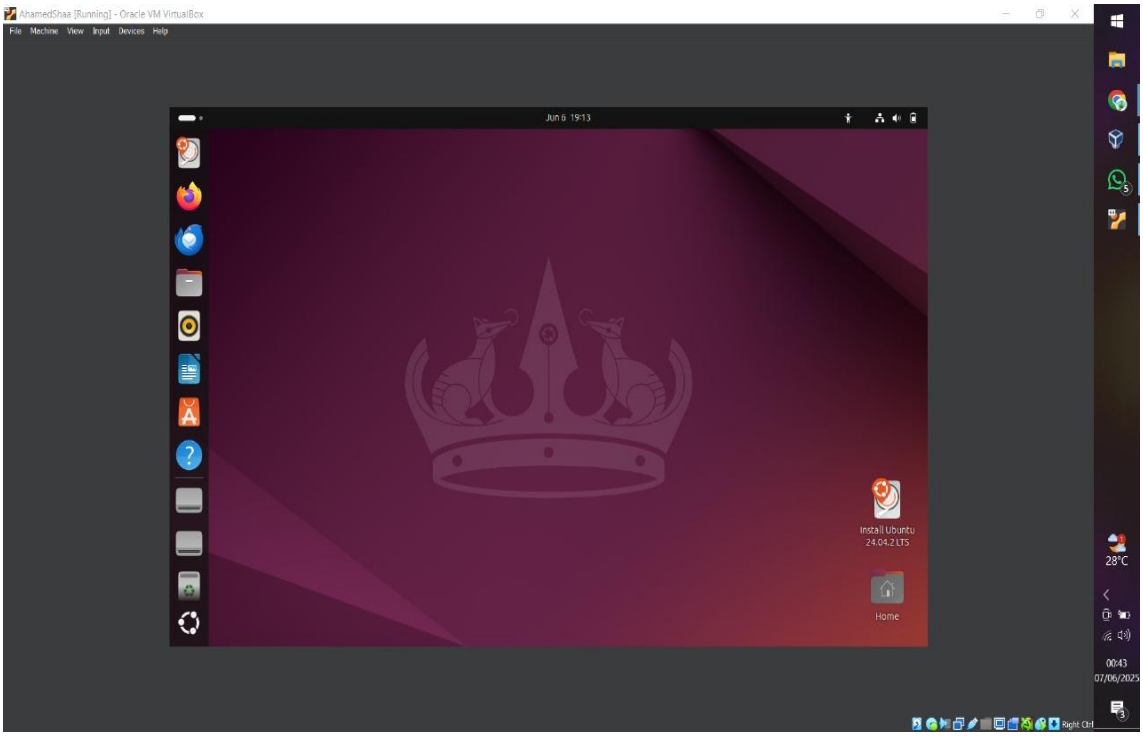
Current results are accurate for simple polarity detection using a lexicon-based approach.

- In order to gain more granular analysis neural sentiment detection can be incorporated.
- Machine learning-based sentiment classifiers like Naïve Bayes and BERT will may give greater accuracy.
- Extend the analysis across multiple months or datasets. This can be used to observe sentiment trends over time.

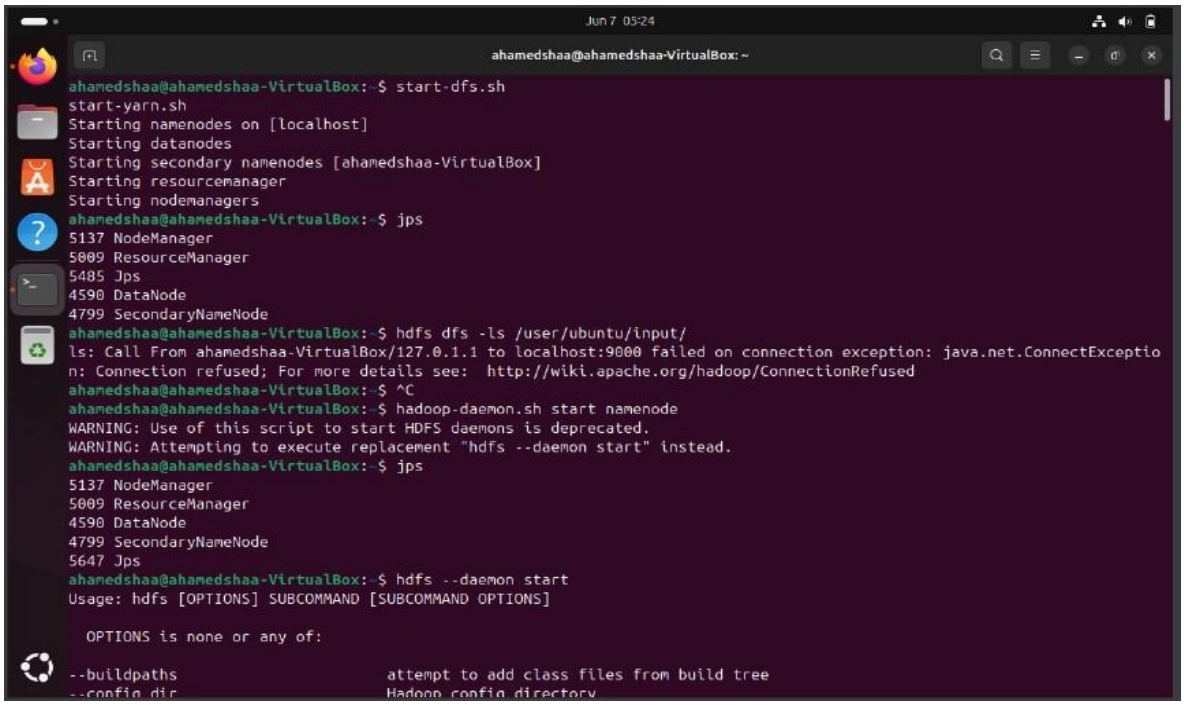
- Future research could broaden the model by analyzing sentiment trends over time, grouping comments by theme using topic modeling, or analyzing engagement by comparing the sentiment of comments to users' interactions or the popularity of articles
- Moving to multi-node Hadoop cluster would enhance performance.

Screenshots

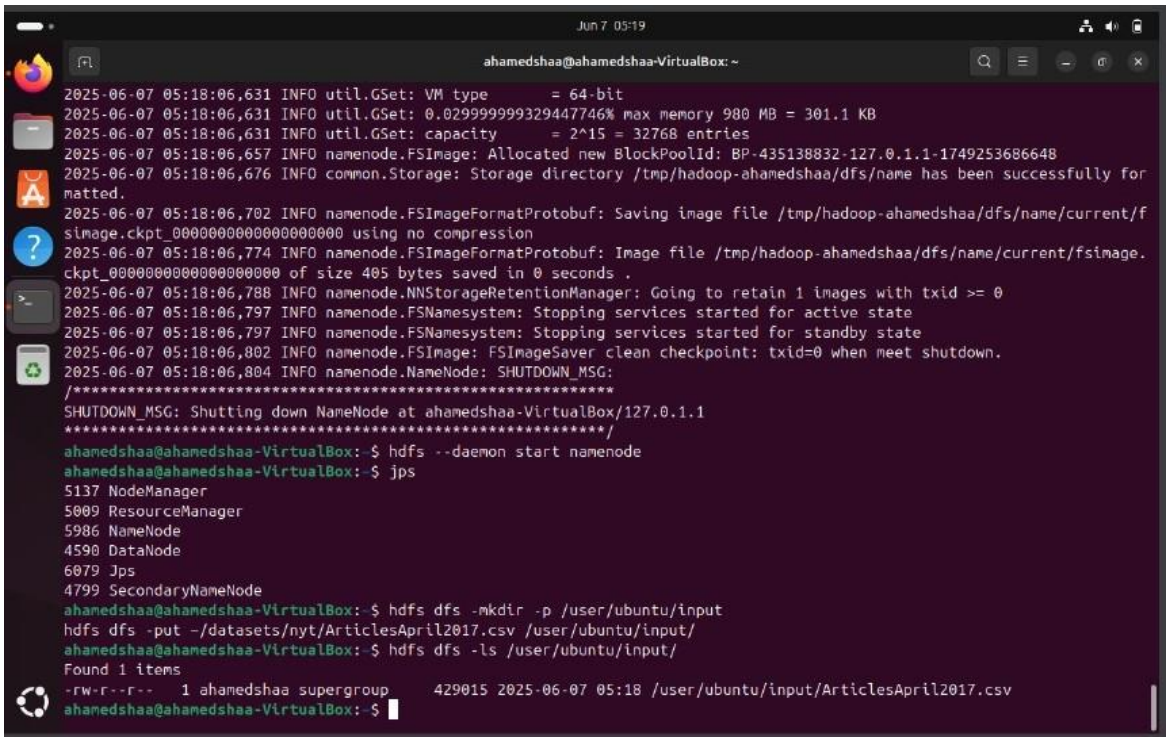
1. Install New ubuntu-24.04.2 in VM Virtual box



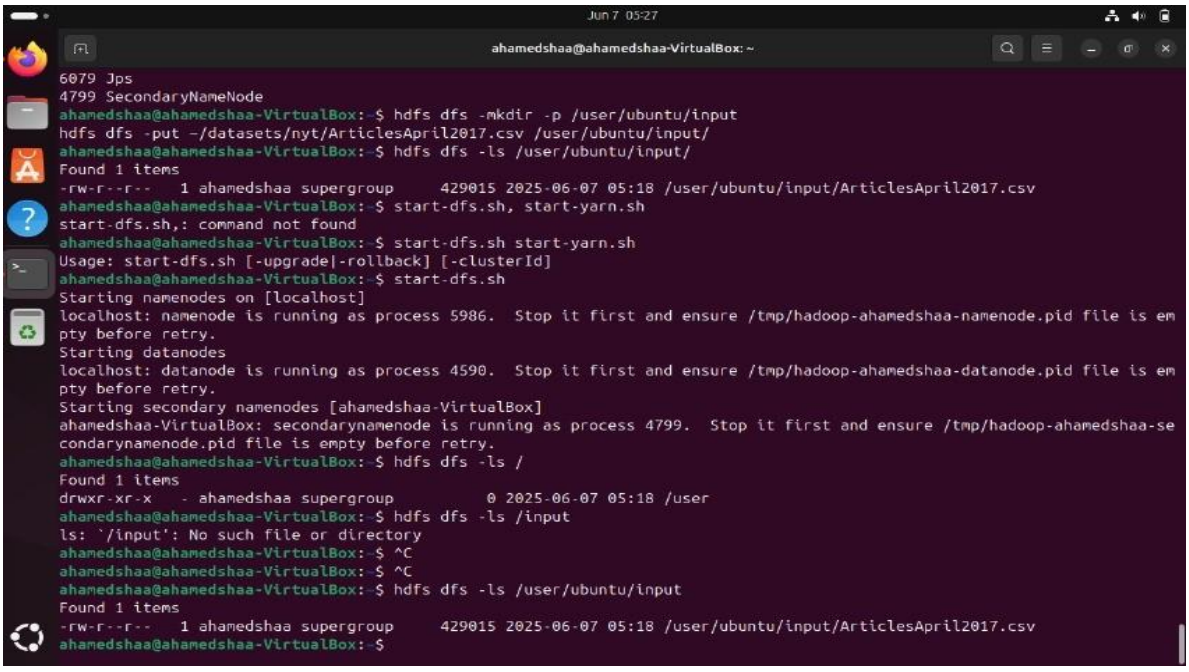
2. HDFS Daemons Started via start-dfs.sh



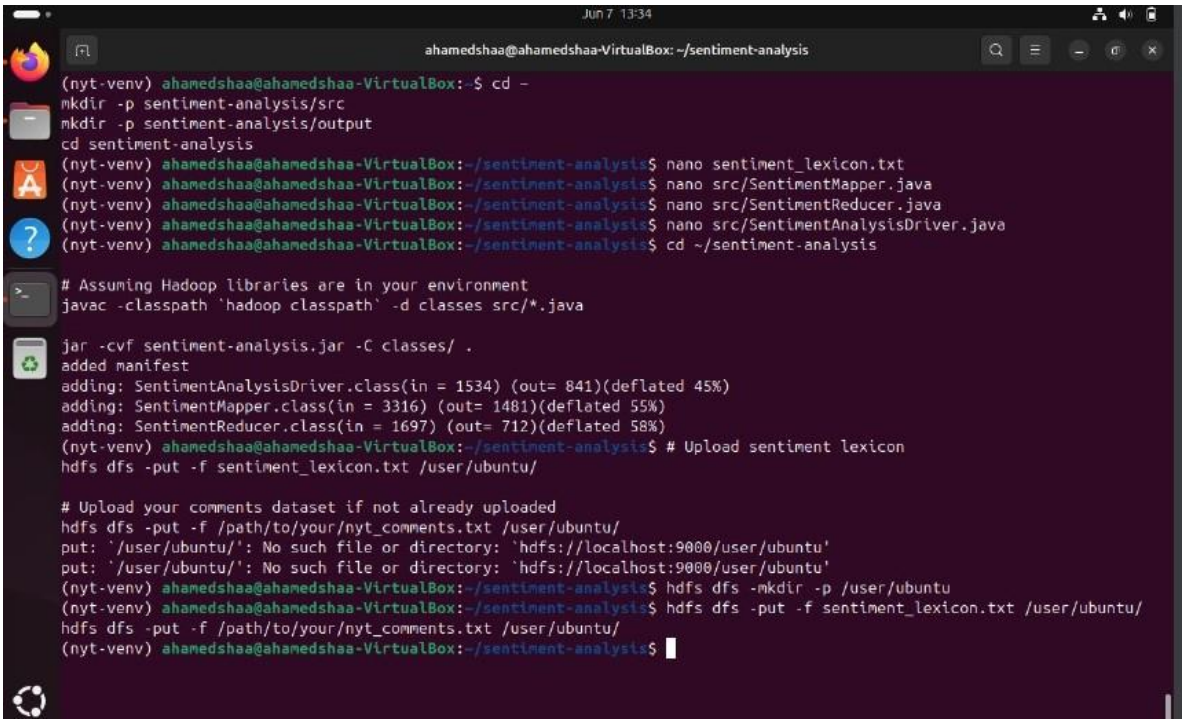
3. Starting NameNode and Verifying Daemons with jps



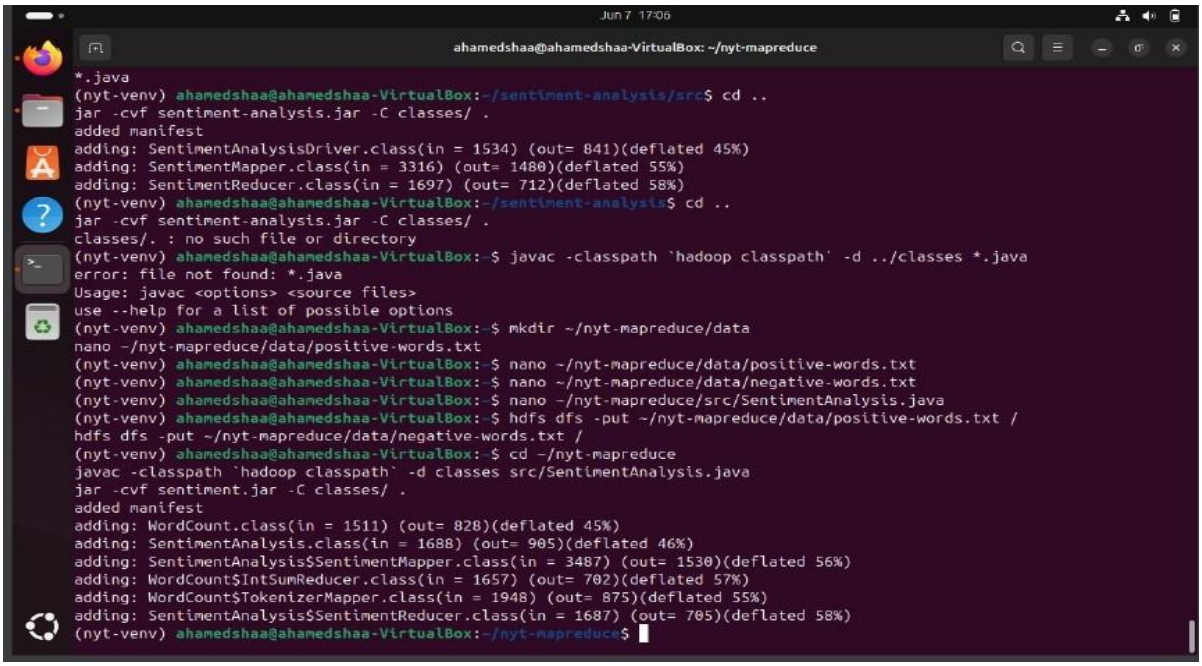
4. Listing Dataset (ArticlesApril2017.csv) in HDFS Using hdfs dfs -ls /input



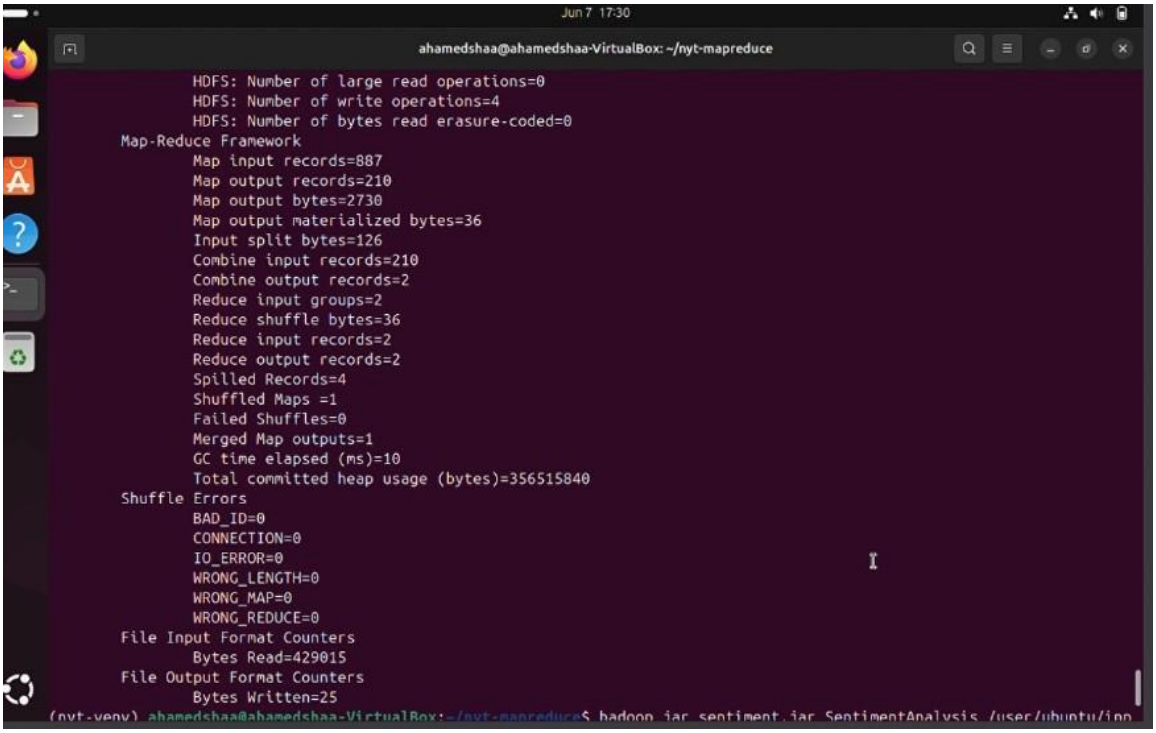
5. Project Directory Structure with Dataset, Lexicon, and Java Source Files



6. Creating Lexicon Files and Compiling SentimentAnalysis.java into Executable JAR for Hadoop



7. MapReduce Job Execution



8. Results plot