

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.

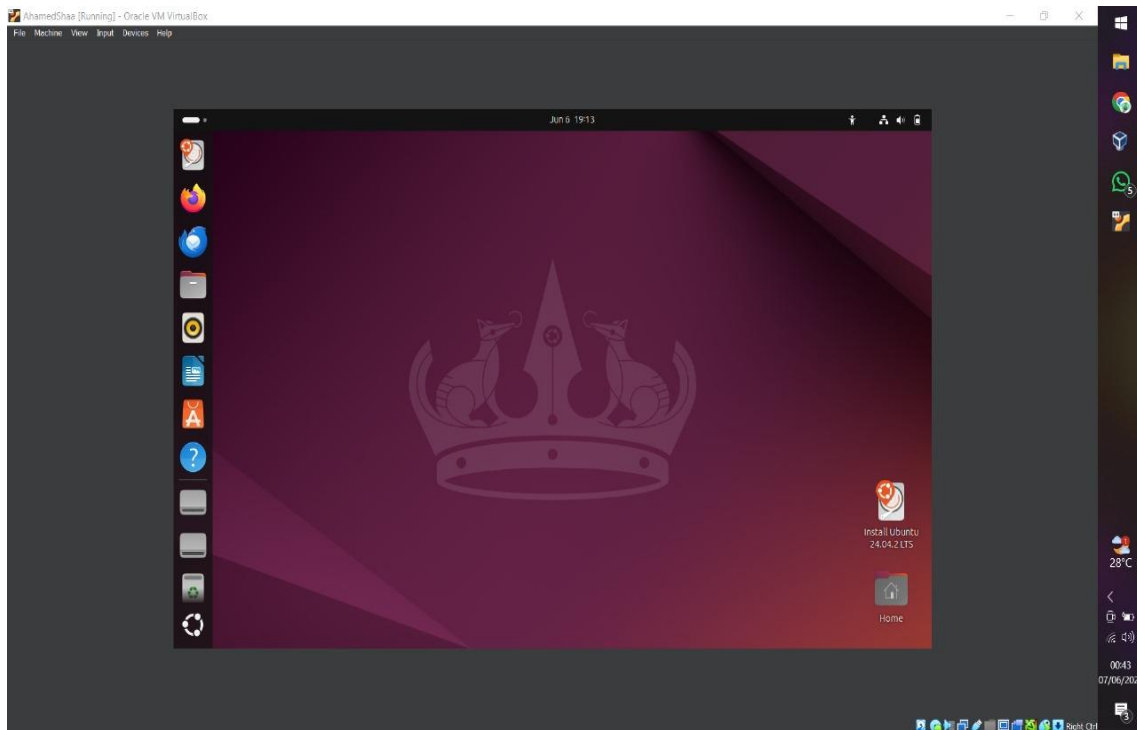
Current results are accurate for simple polarity detection using a lexicon-based approach.

- 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

```
Jun 7 09:24
ahamedshaa@ahamedshaa-VirtualBox: ~
ahamedshaa@ahamedshaa-VirtualBox:~$ start-dfs.sh
start-yarn.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ahamedshaa-VirtualBox]
Starting resourcemanager
Starting nodemanagers
ahamedshaa@ahamedshaa-VirtualBox:~$ jps
5137 NodeManager
5009 ResourceManager
5485 Jps
4590 DataNode
4799 SecondaryNameNode
ahamedshaa@ahamedshaa-VirtualBox:~$ hdfs dfs -ls /user/ubuntu/input/
ls: Call From ahamedshaa-VirtualBox/127.0.1.1 to localhost:9000 failed on connection exception: java.net.ConnectException: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
ahamedshaa@ahamedshaa-VirtualBox:~$ ^C
ahamedshaa@ahamedshaa-VirtualBox:~$ hadoop-daemon.sh start namenode
WARNING: Use of this script to start HDFS daemons is deprecated.
WARNING: Attempting to execute replacement "hdfs --daemon start" instead.
ahamedshaa@ahamedshaa-VirtualBox:~$ jps
5137 NodeManager
5009 ResourceManager
4590 DataNode
4799 SecondaryNameNode
5647 Jps
ahamedshaa@ahamedshaa-VirtualBox:~$ hdfs --daemon start
Usage: hdfs [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]

OPTIONS is none or any of:
--buildpaths          attempt to add class files from build tree
--confdir             Hadoop confdir directory
```

3. Starting NameNode and Verifying Daemons with jps

```
Jun 7 05:19
ahamedshaa@ahamedshaa-VirtualBox: ~
2025-06-07 05:18:06,631 INFO util.GSet: VM type = 64-bit
2025-06-07 05:18:06,631 INFO util.GSet: 0.029999999329447746% max memory 980 MB = 301.1 KB
2025-06-07 05:18:06,631 INFO util.GSet: capacity = 2^15 = 32768 entries
2025-06-07 05:18:06,657 INFO namenode.FSImage: Allocated new BlockPoolId: BP-435138832-127.0.1.1-1749253686648
2025-06-07 05:18:06,676 INFO common.Storage: Storage directory /tmp/hadoop-ahamedshaa/dfs/name has been successfully formatted.
2025-06-07 05:18:06,702 INFO namenode.FSImageFormatProtobuf: Saving image file /tmp/hadoop-ahamedshaa/dfs/name/current/fsimage.ckpt_000000000000000000 using no compression
2025-06-07 05:18:06,774 INFO namenode.FSImageFormatProtobuf: Image file /tmp/hadoop-ahamedshaa/dfs/name/current/fsimage.ckpt_000000000000000000 of size 405 bytes saved in 0 seconds.
2025-06-07 05:18:06,788 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2025-06-07 05:18:06,797 INFO namenode.FSNamesystem: Stopping services started for active state
2025-06-07 05:18:06,797 INFO namenode.FSNamesystem: Stopping services started for standby state
2025-06-07 05:18:06,802 INFO namenode.FSImageSaver: clean checkpoint: txid=0 when meet shutdown.
2025-06-07 05:18:06,804 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at ahamedshaa-VirtualBox/127.0.1.1
*****/
ahamedshaa@ahamedshaa-VirtualBox: $ hdfs --daemon start namenode
ahamedshaa@ahamedshaa-VirtualBox: $ jps
5137 NodeManager
5009 ResourceManager
5986 NameNode
4590 DataNode
6079 Jps
4799 SecondaryNameNode
ahamedshaa@ahamedshaa-VirtualBox: $ hdfs dfs -mkdir -p /user/ubuntu/input
hdfs dfs -put -/datasets/nyt/ArticlesApril2017.csv /user/ubuntu/input/
ahamedshaa@ahamedshaa-VirtualBox: $ hdfs dfs -ls /user/ubuntu/input/
Found 1 items
-rw-r--r-- 1 ahamedshaa supergroup 429015 2025-06-07 05:18 /user/ubuntu/input/ArticlesApril2017.csv
ahamedshaa@ahamedshaa-VirtualBox: $
```

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

```
Jun 7 05:27
ahamedshaa@ahamedshaa-VirtualBox: ~
6079 Jps
4799 SecondaryNameNode
ahamedshaa@ahamedshaa-VirtualBox: $ hdfs dfs -mkdir -p /user/ubuntu/input
hdfs dfs -put -/datasets/nyt/ArticlesApril2017.csv /user/ubuntu/input/
ahamedshaa@ahamedshaa-VirtualBox: $ hdfs dfs -ls /user/ubuntu/input/
Found 1 items
-rw-r--r-- 1 ahamedshaa supergroup 429015 2025-06-07 05:18 /user/ubuntu/input/ArticlesApril2017.csv
ahamedshaa@ahamedshaa-VirtualBox: $ start-dfs.sh, start-yarn.sh
start-dfs.sh,: command not found
ahamedshaa@ahamedshaa-VirtualBox: $ start-dfs.sh start-yarn.sh
Usage: start-dfs.sh [-upgrade|-rollback] [-clusterId]
ahamedshaa@ahamedshaa-VirtualBox: $ start-dfs.sh
Starting namenodes on [localhost]
localhost: namenode is running as process 5986. Stop it first and ensure /tmp/hadoop-ahamedshaa-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 4590. Stop it first and ensure /tmp/hadoop-ahamedshaa-datanode.pid file is empty before retry.
Starting secondary namenodes [ahamedshaa-VirtualBox]
ahamedshaa-VirtualBox: secondarynamenode is running as process 4799. Stop it first and ensure /tmp/hadoop-ahamedshaa-secondarynamenode.pid file is empty before retry.
ahamedshaa@ahamedshaa-VirtualBox: $ hdfs dfs -ls /
Found 1 items
drwxr-xr-x 1 ahamedshaa supergroup 0 2025-06-07 05:18 /user
ahamedshaa@ahamedshaa-VirtualBox: $ hdfs dfs -ls /input
ls: '/input': No such file or directory
ahamedshaa@ahamedshaa-VirtualBox: $ ^C
ahamedshaa@ahamedshaa-VirtualBox: $ ^C
ahamedshaa@ahamedshaa-VirtualBox: $ hdfs dfs -ls /user/ubuntu/input
Found 1 items
-rw-r--r-- 1 ahamedshaa supergroup 429015 2025-06-07 05:18 /user/ubuntu/input/ArticlesApril2017.csv
ahamedshaa@ahamedshaa-VirtualBox: $
```


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

```
Jun 7 13:34
ahamedshaa@ahamedshaa-VirtualBox: ~/sentiment-analysis

(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~$ cd -
mkdir -p sentiment-analysis/src
mkdir -p sentiment-analysis/output
cd sentiment-analysis
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~/sentiment-analysis$ nano sentiment_lexicon.txt
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~/sentiment-analysis$ nano src/SentimentMapper.java
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~/sentiment-analysis$ nano src/SentimentReducer.java
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~/sentiment-analysis$ nano src/SentimentAnalysisDriver.java
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~/sentiment-analysis$ cd ~/sentiment-analysis

# Assuming Hadoop libraries are in your environment
javac -classpath 'hadoop classpath' -d classes src/*.java

jar -cvf sentiment-analysis.jar -C classes/ .
added manifest
adding: SentimentAnalysisDriver.class(in = 1534) (out= 841)(deflated 45%)
adding: SentimentMapper.class(in = 3316) (out= 1481)(deflated 55%)
adding: SentimentReducer.class(in = 1697) (out= 712)(deflated 58%)
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~/sentiment-analysis$ # Upload sentiment lexicon
hdfs dfs -put -f sentiment_lexicon.txt /user/ubuntu/

# Upload your comments dataset if not already uploaded
hdfs dfs -put -f /path/to/your/nyt_comments.txt /user/ubuntu/
put: '/user/ubuntu/': No such file or directory: 'hdfs://localhost:9000/user/ubuntu'
put: '/user/ubuntu/': No such file or directory: 'hdfs://localhost:9000/user/ubuntu'
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~/sentiment-analysis$ hdfs dfs -mkdir -p /user/ubuntu
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~/sentiment-analysis$ hdfs dfs -put -f sentiment_lexicon.txt /user/ubuntu/
hdfs dfs -put -f /path/to/your/nyt_comments.txt /user/ubuntu/
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~/sentiment-analysis$
```

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

```
Jun 7 17:06
ahamedshaa@ahamedshaa-VirtualBox: ~/nyt-mapreduce

*.java
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~/sentiment-analysis/src$ cd ..
jar -cvf sentiment-analysis.jar -C classes/ .
added manifest
adding: SentimentAnalysisDriver.class(in = 1534) (out= 841)(deflated 45%)
adding: SentimentMapper.class(in = 3316) (out= 1480)(deflated 55%)
adding: SentimentReducer.class(in = 1697) (out= 712)(deflated 58%)
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~/sentiment-analysis$ cd ..
jar -cvf sentiment-analysis.jar -C classes/ .
classes/. : no such file or directory
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~$ javac -classpath 'hadoop classpath' -d ../classes *.java
error: file not found: *.java
Usage: javac <options> <source files>
use --help for a list of possible options
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~$ mkdir ~/nyt-mapreduce/data
nano ~/nyt-mapreduce/data/positive-words.txt
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~$ nano ~/nyt-mapreduce/data/positive-words.txt
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~$ nano ~/nyt-mapreduce/data/negative-words.txt
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~$ nano ~/nyt-mapreduce/src/SentimentAnalysis.java
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~$ hdfs dfs -put ~/nyt-mapreduce/data/positive-words.txt /
hdfs dfs -put ~/nyt-mapreduce/data/negative-words.txt /
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~$ cd ~/nyt-mapreduce
javac -classpath 'hadoop classpath' -d classes src/SentimentAnalysis.java
jar -cvf sentiment.jar -C classes/ .
added manifest
adding: WordCount.class(in = 1511) (out= 828)(deflated 45%)
adding: SentimentAnalysis.class(in = 1688) (out= 905)(deflated 46%)
adding: SentimentAnalysis$SentimentMapper.class(in = 3487) (out= 1530)(deflated 56%)
adding: WordCount$IntSumReducer.class(in = 1657) (out= 702)(deflated 57%)
adding: WordCount$TokenizerMapper.class(in = 1948) (out= 875)(deflated 55%)
adding: SentimentAnalysis$SentimentReducer.class(in = 1687) (out= 705)(deflated 58%)
(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox:~/nyt-mapreduce$
```

7. MapReduce Job Execution

```
Jun 7 17:30
ahamedshaa@ahamedshaa-VirtualBox: ~/nyt-mapreduce

HDFS: Number of large read operations=0
HDFS: Number of write operations=4
HDFS: Number of bytes read erasure-coded=0

Map-Reduce Framework
  Map input records=887
  Map output records=210
  Map output bytes=2730
  Map output materialized bytes=36
  Input split bytes=126
  Combine input records=210
  Combine output records=2
  Reduce input groups=2
  Reduce shuffle bytes=36
  Reduce input records=2
  Reduce output records=2
  Spilled Records=4
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=10
  Total committed heap usage (bytes)=356515840

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0

File Input Format Counters
  Bytes Read=429015
File Output Format Counters
  Bytes Written=25

(nyt-venv) ahamedshaa@ahamedshaa-VirtualBox: ~/nyt-mapreduce$ hadoop jar sentiment.jar SentimentAnalysis /user/ubuntu/ion
```

8. Results plot

