iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.37193

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Hadoop Based Generic Template for Performing Sentiment Analysis Using Apache PIG

Gadige Vishal Sai[1], Chukka Nikhil[2], Panyala Harsha Vardhan Reddy[3], Sai Harsha Bandarupally[4]
[1]*Software Development Engineer, NCR Corporation, Hyderabad*
[2]*Application Developer, Accenture, Hyderabad*
[3]*Associate Software Engineer, CGI, Hyderabad*
[4]*Full Stack Engineering Analyst, Accenture, Hyderabad*

Abstract: ***Every day over 2.5 quintillion data is generated using various channels like online surveys, transactional data tracking, social media monitoring, etc. Out of these majority of the data is generated using social media platforms. This raw data contains information that can be used for industrial, economic, social and business purposes. To facilitate this, sentiment analysis has become a prospect for various tech-based industry giants to review and analyze their products. Hadoop has been established as one of the best tools for storing, processing, and streaming data in the market. In this paper, we present a generic approach to performing sentiment analysis using Apache PIG which classifies the given data taken from a dataset to either positive or negative to get the people's sentiment over an object or an issue.***

Keywords: Big Data Analysis, Hadoop, Apache PIG, Sentiment Analysis, Opinion Mining, Business Analysis

I. INTRODUCTION

Since the dawn of time up until 2005, humans have created nearly 130 exabytes of data. Nevertheless, this value has been significantly increased to 9,100 and later to 40,900 during 2005 to 2015 and 2015 to 2020. This became a massive challenge for the World Wide Web to understand and satisfy the requirements of people. At present, people are comprehensively conveying their thoughts over discussion forms, online blogs and social networking applications. Industrial organizations have understood this and applied various techniques to study the people's sentiment over their products, which help them improve their productivity by predicting their products' success rate. However, to perform this, large scale data storing and processing needs to be done. To answer this, Hadoop has been recommended as one of the best tools for data storing and processing techniques.

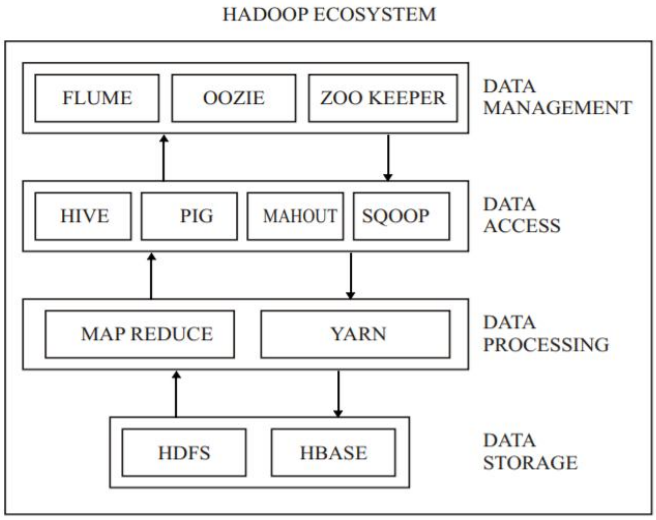


Figure 1. Hadoop Ecosystem

The above figure clearly illustrates the various tools that are available in the Hadoop ecosystem. Hadoop consists of various storing and processing components that can handle all forms of data – structured, semi-structured, quasi structured, and unstructured. In this paper, we use Hadoop's processing component Pig to create a generic template to perform sentiment analysis on a pre-processed data set.

©IJRASET: All Rights are Reserved

II. RELATED WORK

Big Data analytics is the investigative process that includes recording, storing, analyzing, unsheathing helpful information from the data. It uses different scientific methods, algorithms and techniques to analyze the data effectively and efficiently; data can be either structured or unstructured. Many industries like banking, finance, e-commerce, manufacturing, etc. use data science to manage gigantic data from multiple sources to derive valuable information and make smarter data-driven decisions.

Ajit Noonia and team performed sentimental analysis on Twitter using PIG and HIVE [1], where they collected over 5000 tweets, performing pre-processing over it. They performed sentimental analysis so that they can differentiate between positive and negative tweets. In this paper, researchers have changed the data from unstructured data to structured data to perform pre-processing. After pre-processing, they used tokenization to convert text into tokens before transforming it, and data comparison is made using a data dictionary.

Another related work includes Opinion Mining of Twitter Data using Hadoop and Apache Pig [2]. It proposes a method of using a dictionary-based approach for opinion mining. With the help of pig statements, they have analyzed Twitter data to check the tweets' polarity based on polarity dictionary to separate tweets into positive and negative tweets (opinions). This paper has analyzed the Twitter data using some visualizing techniques to analyze the results effectively.

III.METHODOLOGY

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

In this paper, we propose a generic template which facilitates the user to perform sentiment analysis by implementing the following methodology.

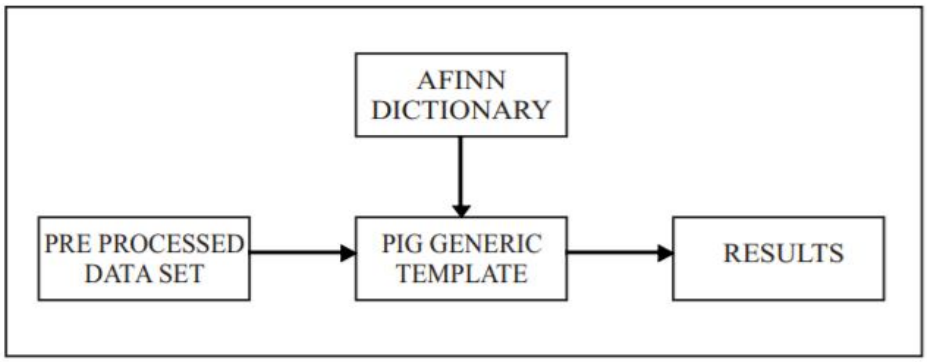


Figure 2. The above diagram illustrates the proposed methodology

A. *Preprocessed Dataset*

In this proposed experiment, we perform sentiment analysis on a preprocessed dataset. The process of performing preprocessing is outside of the scope of this experiment. It should be done in such a way that all the types of data – structured, semi-structured, quasi structured, and unstructured should be preprocessed to a set of English language statements. This set of sentences should be fed as input to our proposed template, where it helps the user to perform sentiment analysis along with the help of the AFINN dictionary.

B. *AFINN Dictionary*

AFINN dictionary is an open-source distributed lexicon that contains a list of English words, each associated with a corresponding integer value. Finn Arup Nielsen developed it. The dictionary consists of 2,477 coded words, each having a value score between –5 and +5. A negative score implies that the word is associated negatively, and an upbeat score implies that the word is associated with a positive context. By mapping the word score values to the input statements, we perform sentiment analysis.

```
absorbed    1
abuse  -3
abused      -3
abuses-3
abusive     -3
accept 1
accepted    1
```

Figure 3. Screenshot containing data in AFINN dictionary

©IJRASET: All Rights are Reserved

C. Proposed PIG Generic Template

The sentiment analysis template consists of a series of PIG commands that uses both the input set of statements and the AFINN dictionary. The sequence of statements executed are shown below using the flowchart:

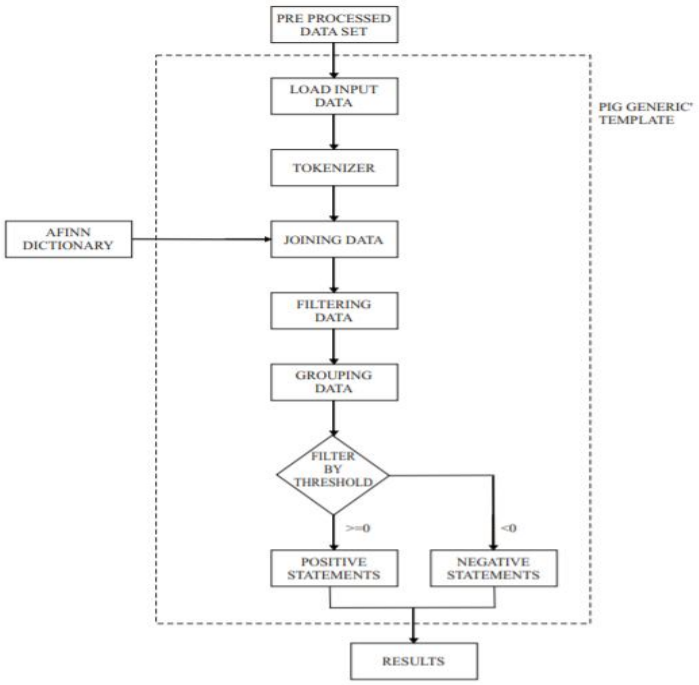


Figure 4. The above figure illustrates a detailed step by step implementation of the proposed generic template

Generally, PIG applications can be executed in two modes – local mode and HDFS (Hadoop Distributed File System) mode. Usually, the local mode is used for software development, whereas the HDFS mode is used for performing software testing. Here, we use the local mode for executing our PIG script. All PIG applications are written in PIG Latin language that can be compiled in the grunt shell.

Here, we used the dataset related to the people's opinion on demonetization in India for testing the application.

1) *Load Input Data:* Here we use the PIG Latin's load command to load the data onto the PIG local system. Using the PigStorage() function, we perform the load operation from a given local path. We can use the dump command to view the loaded data, which prints the loaded data on the grunt shell. The dump command is like a select query in RDBMS. We can also use the describe command, which helps us view the structure of the requested data.

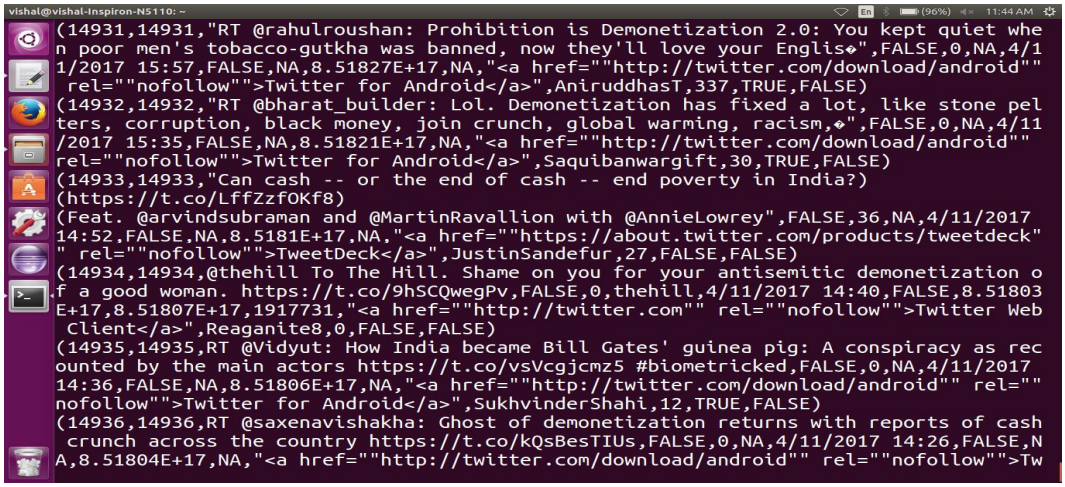


Figure 5. Screenshot demonstrating the load command

©IJRASET: All Rights are Reserved

2) *Tokenize the Data:* A sentence is a collection of words. The meaning of the words combined represents the entire meaning of the sentence. To know the sentence's sentiment, we break down each sentence into tokens of the format word and word count. Later we use the flatten command to apply it to all the available sentences. The structure of the data looks as follows:

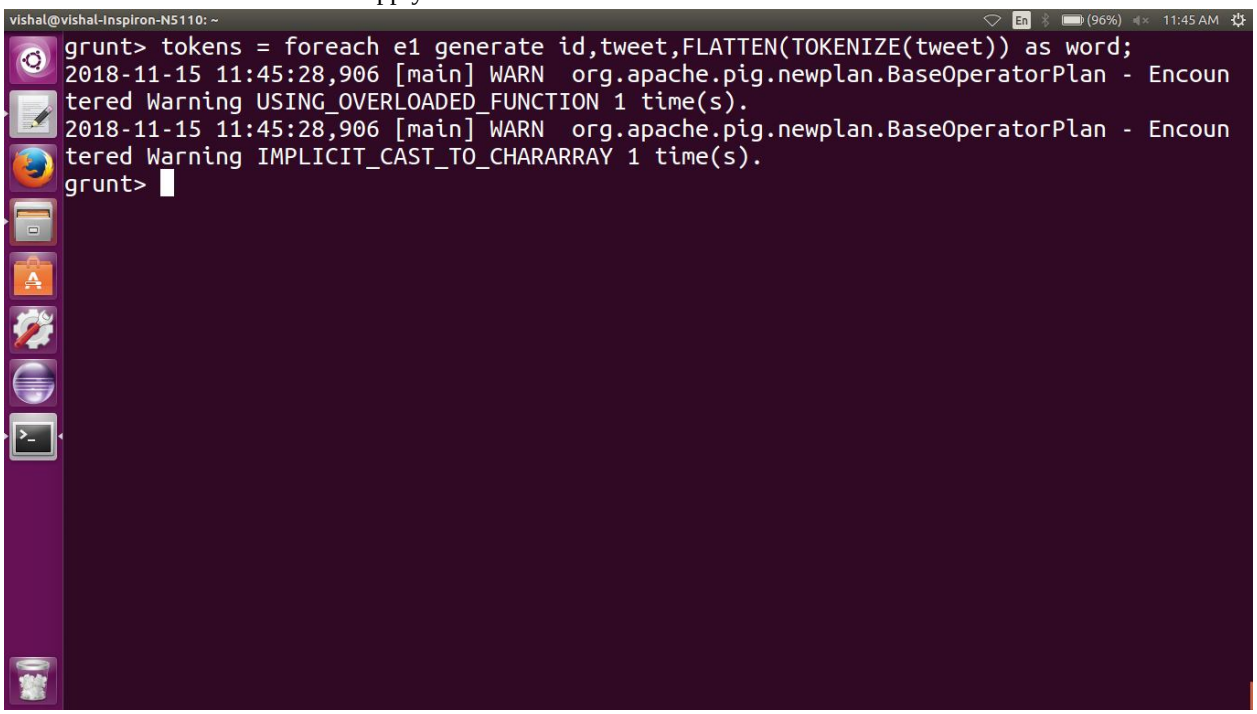


Figure 6. Screenshot demonstrating the process of tokenization

3) *Load AFINN dictionary:* We use the same load command to load the AFINN dictionary to PIG local system. Also, to view the structure of the data, we can use the describe command.

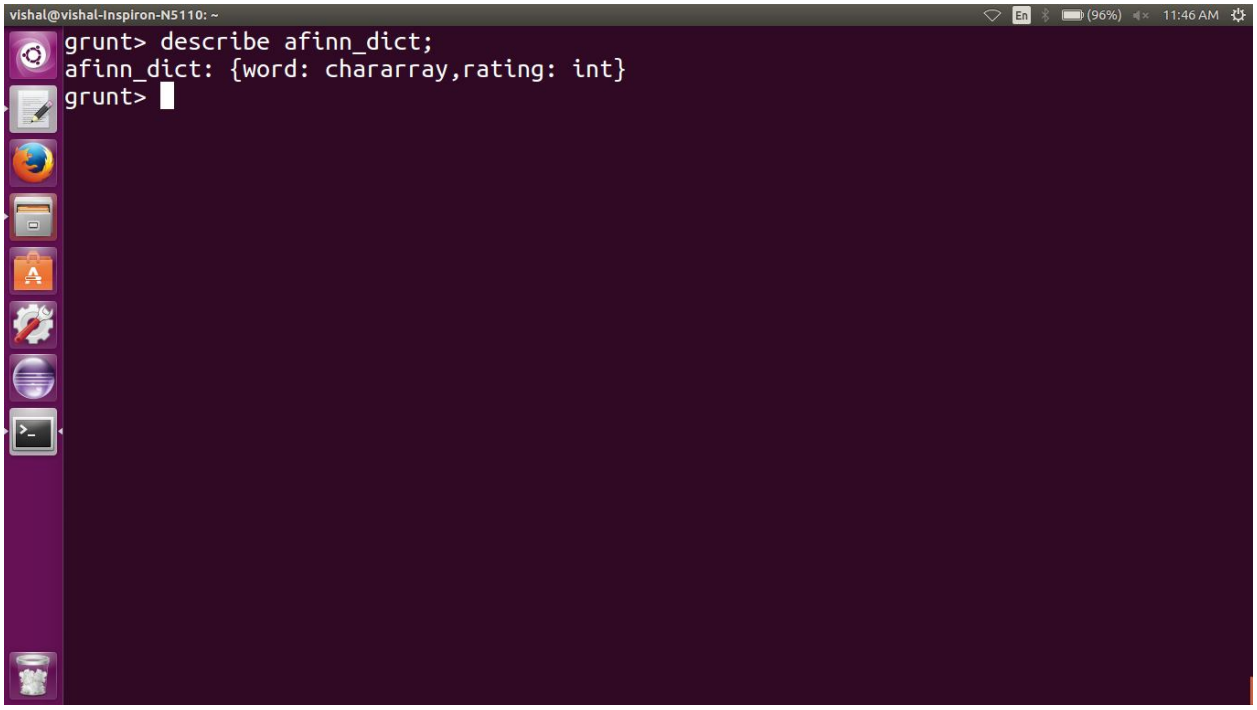


Figure 7. Screenshot demonstrating the structure of AFINN dictionary

©IJRASET: All Rights are Reserved

4) *Joining the Data:* To perform sentiment analysis, we need all the data to merge into a single large chunk of data. To achieve this, we use the concept of joins. Specifically, we use the left outer join as we are more concerned about the words present in both the AFINN dictionary and the data set.



Figure 8. Screenshot showing the join operation on the data

5) *Filtering the Data:* Here we use the filter command, which helps us filter out the data columns essential for the analysis. Here, we can also apply the describe command to view the structure of the filtered output.



Figure 9. Screenshot showing the filtering process

©IJRASET: All Rights are Reserved

6) *Grouping the Data:* In PIG, we need to apply the group command to perform any analysis or aggregate operation.

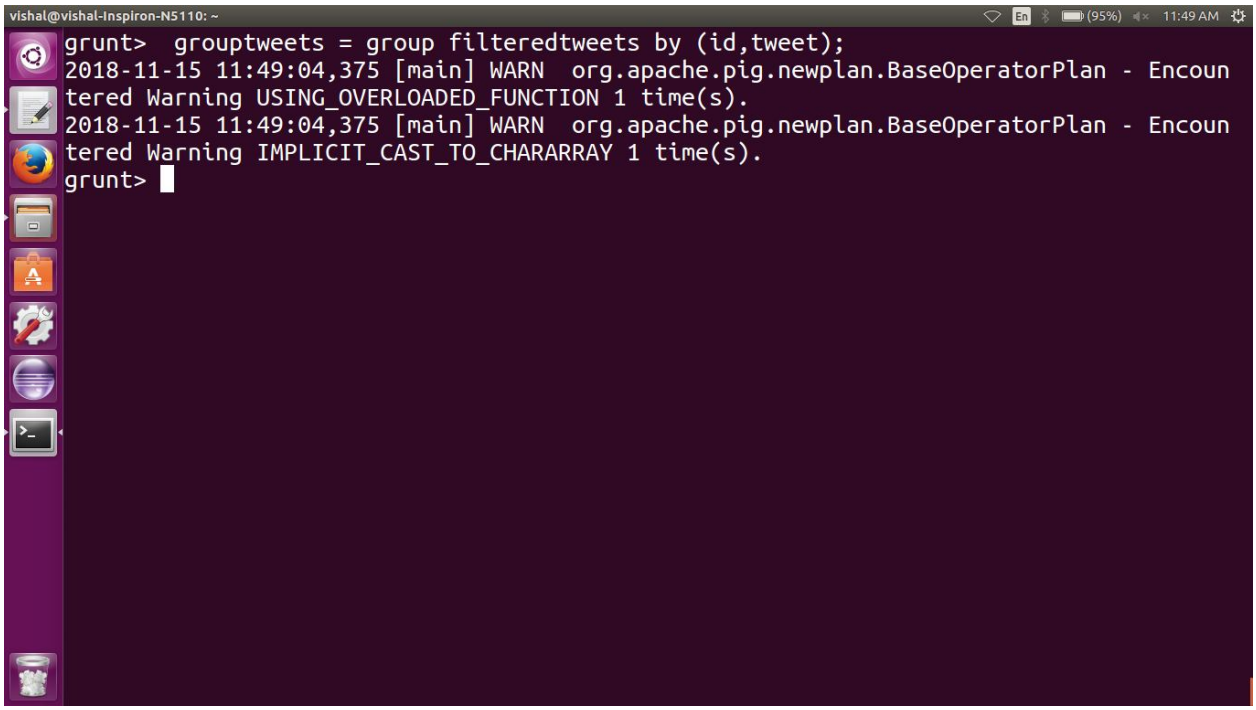


Figure 10. Screenshot showing the execution of group command

7) *Filter by threshold:* The main advantage of the AFINN dictionary is that all the negative words have a negative polarity, and all the positive words have a positive polarity. Here we use the same decision logic to classify data into positive or negative. If the statement's average polarity is greater than or equal to zero, then we say that the statement is positive, else negative.

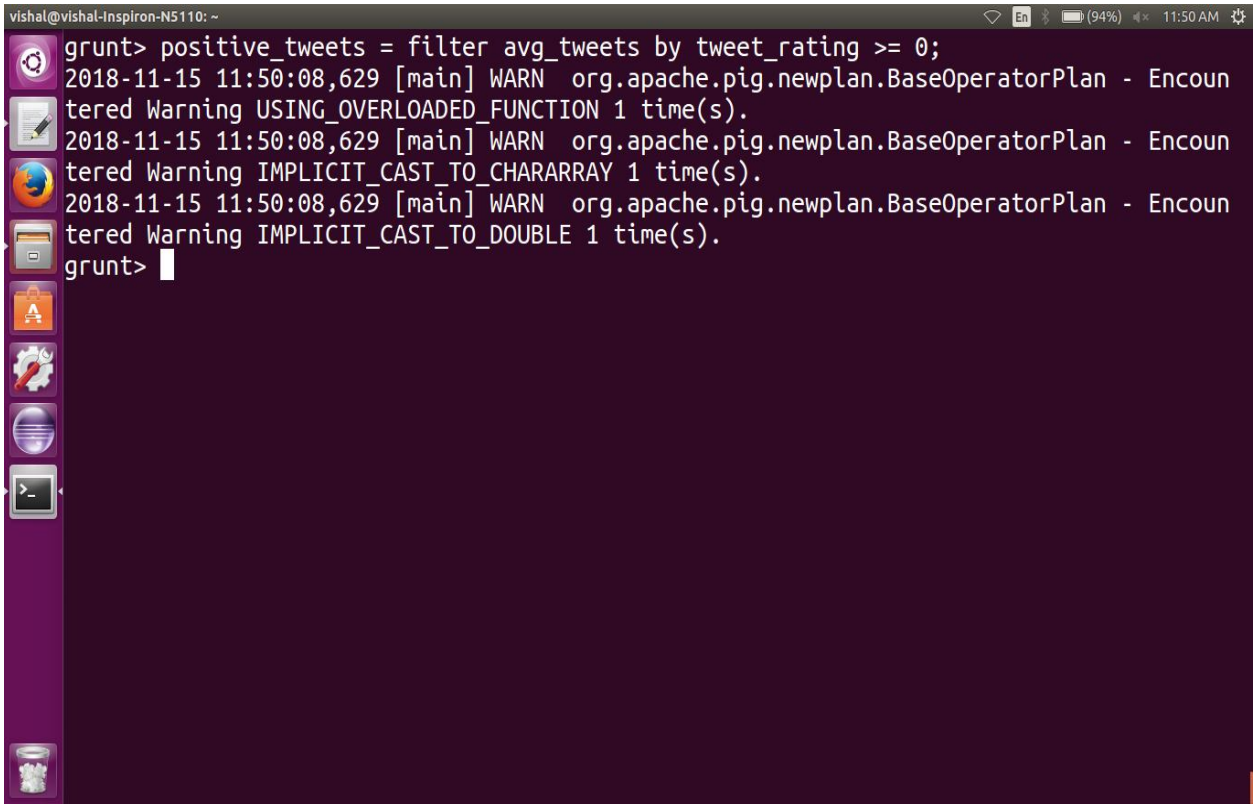


Figure 10. Screenshot showing the logic on how to obtain positive polarity data

©IJRASET: All Rights are Reserved

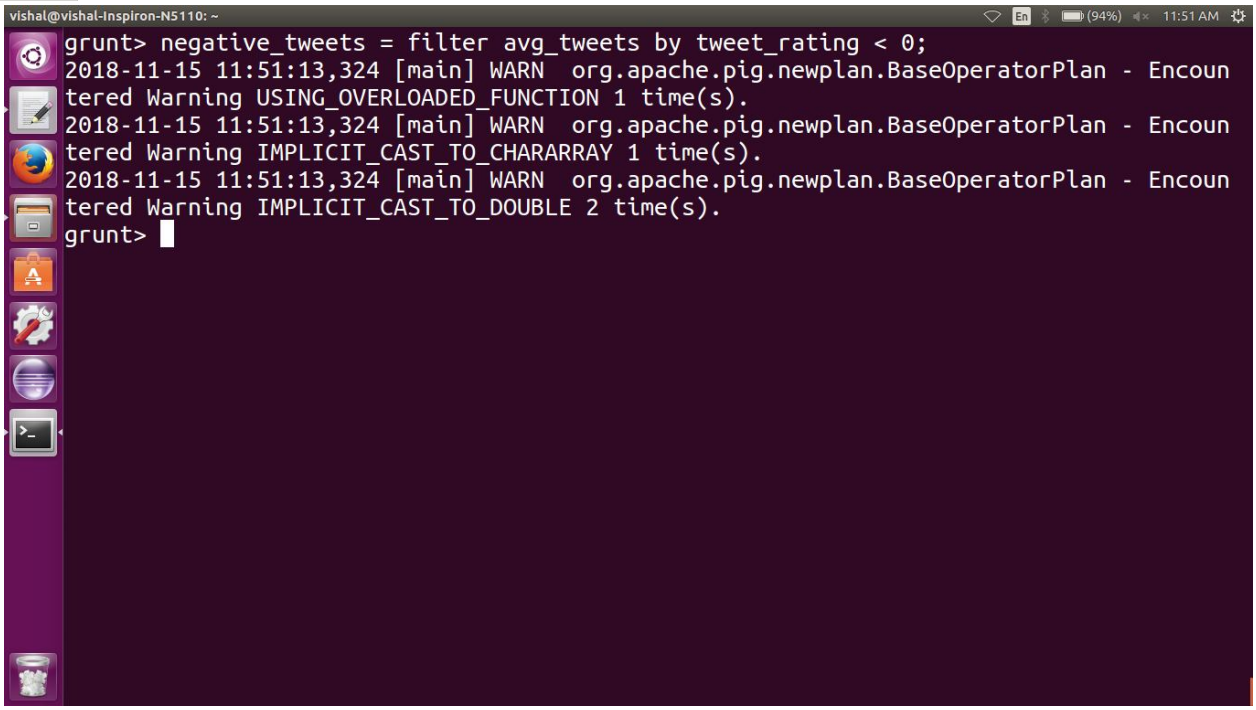


Figure 11. Screenshot showing the logic on how to obtain negative polarity data

We collect all the positive sentences and negative sentences and send them as the output, which can be collected in the subsequent phases. This data can be used in business analytics, data visualization, progress tracking, and other various applications.

IV. RESULTS AND DISCUSSIONS

The generic sentiment analysis template takes the input as processed data statements and produces positive and negatively polarized statements. After fetching the results from the generic sentiment analysis template, one can either use that data for business analysis or data visualization, mainly for predicting the success rate of a product.

Here in this experiment, we considered the Twitter data set containing people's opinion on demonetization in India. The dataset consisted of 14,941 tweets with 16 attributes. However, after performing sentiment analysis using the proposed template, one can see the results as follows:

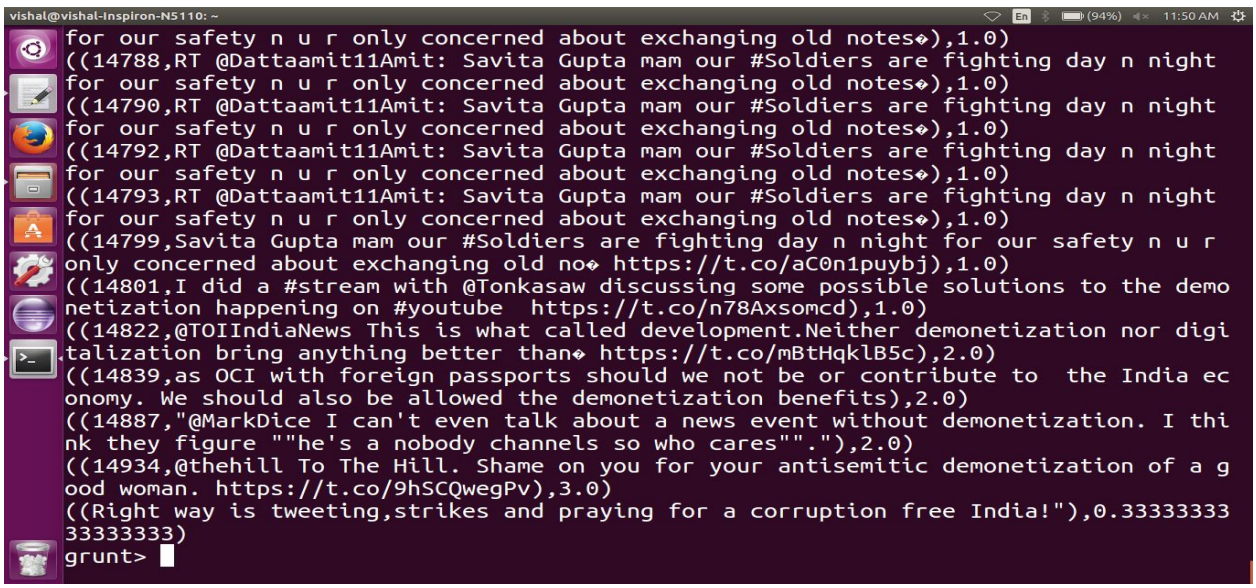


Figure 12. Screenshot showing the positive tweets

©IJRASET: All Rights are Reserved



Figure 13. Screenshot showing the negative tweets

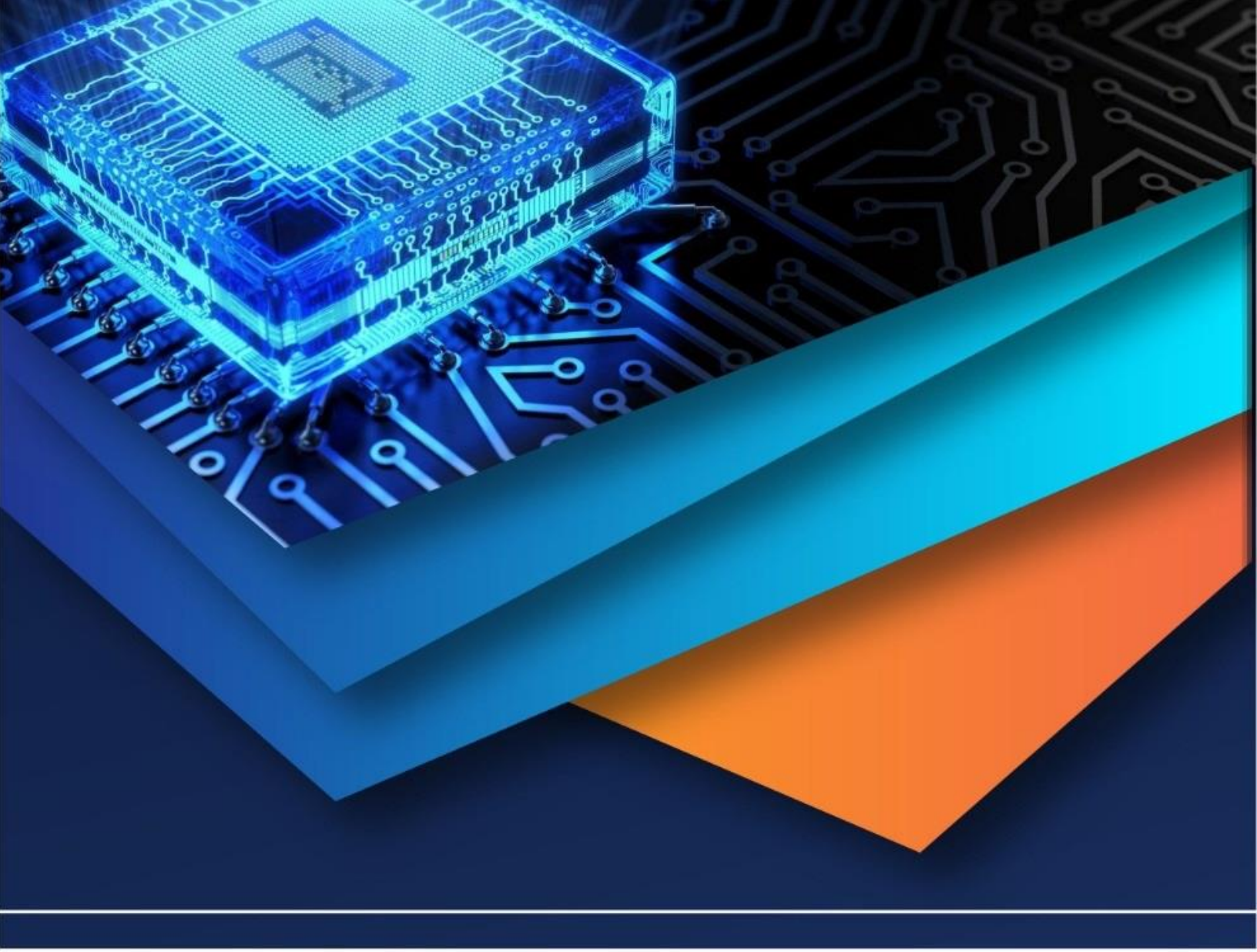
V. CONCLUSION AND FUTURE SCOPE

Hadoop has been effectively proved as one of the best tools for supporting processing, storage, and streaming data. With a vast arsenal of open source software and tools, one can use them according to the demands as specified by the application. PIG facilitates processing the data to obtain favourable outcomes, which can be helpful in the future. In this paper, we created a generic template that helps the businesses understand their product reviews, how clients view their brand and considering public opinion to make business decisions more customer-centric. The project's future scope could be performing sentiment analysis on live data, which can be captured by using either a third-party application or APIs provided by the data vendors themselves. The project's scalability could also be improved if the PIG environment is run on HDFS mode rather than the local system mode.

REFERENCES

[1] Dr. Ajit Noonia (2020), “Sentiment Analysis on Twitter Using Pig and Hive”, International Journal of Innovative Technology and Exploring Engineering, ISSN: 2278- 3075, Volume 9 Issue 3, January 2020W.
[2] Anjali Barskar and Ajay Phulre (2017), “Opinion Mining of Twitter Data using Hadoop and Apache PIG”, International Journal of Computer Applications (0975-8887), Volume 158-No 9, January 2017
[3] Kharde and Sonawane (2016), “Sentimental Analysis of twitter data- A survey of techniques”, International Journal of Computer Applications, 6(10), pp. 2321-9637
[4] Kumar and Sebestian (2015), “Sentiment Analysis on Twitter”, IJCSI International Journal of Computer Science.
[5] Dhawan and Rathee (2013), “Big Data Analytics using Hadoop Components like Pig and Hive”, American International Journal of Research in Science, Technology, Engineering & Mathematics, pp. 88-93.
[6] Shubham Goyal (2016), “Sentimental Analysis of Twitter Data using Text Mining and Hybrid Classification Approach”, International Journal of Advance Research, Ideas and Innovations in Technology, 2(5) , pp. 2454-132XJ.
[7] Rahul Kumar Chawda, Dr. Ghanshyam Thakur, “Big Data and Advanced Analytics Tools”, 2016 Symposium on Colossal Data Analysis and Networking (CDAN), IEEE 2016, ISSN: 978-1-5090-0669-4/16.
[8] Mahalakshmi R, Suseela S , “Big-SoSA:Social Sentiment Analysis and Data Visualization on Big Data”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 4, April 2015 , pp 304-306, ISSN : 2278-1021.
[9] Manoj Kumar Danthala, “Tweet Analysis: Twitter Data processing Using Apache Hadoop”, International Journal Of Core Engineering & Management (IJCEM) Volume 1, Issue 11, February 2015, pp 94-102.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)