iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41675

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: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

Twespa: A Modern Approach to Twitter Sentiment Analysis

Isha Rane[1], Sagarika Sardesai[2], Rucha Tallur[3]
[1, 2, 3]*Computer Engineering, Mumbai University*

Abstract: ***Sentiment analysis has become a stimulating field for both research and industrial domains. The expression sentiment refers to the emotions or thought of the person across some certain issues. Furthermore, it's also considered an immediate application for opinion mining. the large amount of tweets jotted down daily makes Twitter an upscale source of textual data and one among the foremost essential data volumes; therefore, this data has different aims, like business, industrial or social aims consistent with the info requirement and needed processing. Actually, the quantity of knowledge , which is very large , grows rapidly per second and this is often called big data which needs special processing techniques and high computational power so as to perform the specified mining tasks. during this work, we perform a sentiment analysis with the assistance of PySpark framework, an interface for Apache Spark in Python which is taken into account an open source distributed processing platform which utilizes distributed memory abstraction. The goal of using PySpark is that we can run applications parallelly on the distributed cluster (multiple nodes). The effectiveness of our proposed approach is proved against other approaches achieving better classification results when using Naïve Bayes, Logistic Regression and Decision trees classification algorithms. Finally, our solution estimates the performance of Apache Spark concerning its scalability.***
Keywords: Python, Pyspark, Machine learning, tweepy, tokenization.

I. INTRODUCTION

Now social networks sites make the entire world a small village, where users can share their views, feelings, experiences, advice through those sites so that others can get help from these. Since many of us use social media daily, a huge quantity of comments, opinion, article have been created. locating an automatic manner for investigating and classifying users' opinions in social networks could be quite essential. This is mainly because it is considered a great tool for getting direct notes or information from users. The method of classifying texts or documents in keeping with their polarity is referred to as Sentiment Analysis (SA). Sentiment analysis can be described as a major branch of Natural Language Processing (NLP), its aim to identify the meaning from a document in order to discover the polarity of the text.

For the sentiment analysis, we focus our attention in the direction of the Twitter, a micro-blogging social networking website, where users can communicate with each other or share their opinions in short blogs. Large different number of text posts exists on twitter which increases every day, the rapid enormous data growth make the existing databases unable to handle an extensive amount of data in a short time. Also, these databases type designed to process structured data but there is a limitation on it when dealing with huge data. So the conventional solutions are not helpful for organizations to manage and process unstructured or large data. Frameworks, such as Hadoop, Apache Spark, Apache flume and distributed data storages like Hadoop Distributed File System (HDFS), Cassandra and HBase are being very widespread, as they are designed in a manner which facilitates the process of huge amounts of big data and makes it almost effortless

II. MATERIALS AND METHODS

A. *Authentication*

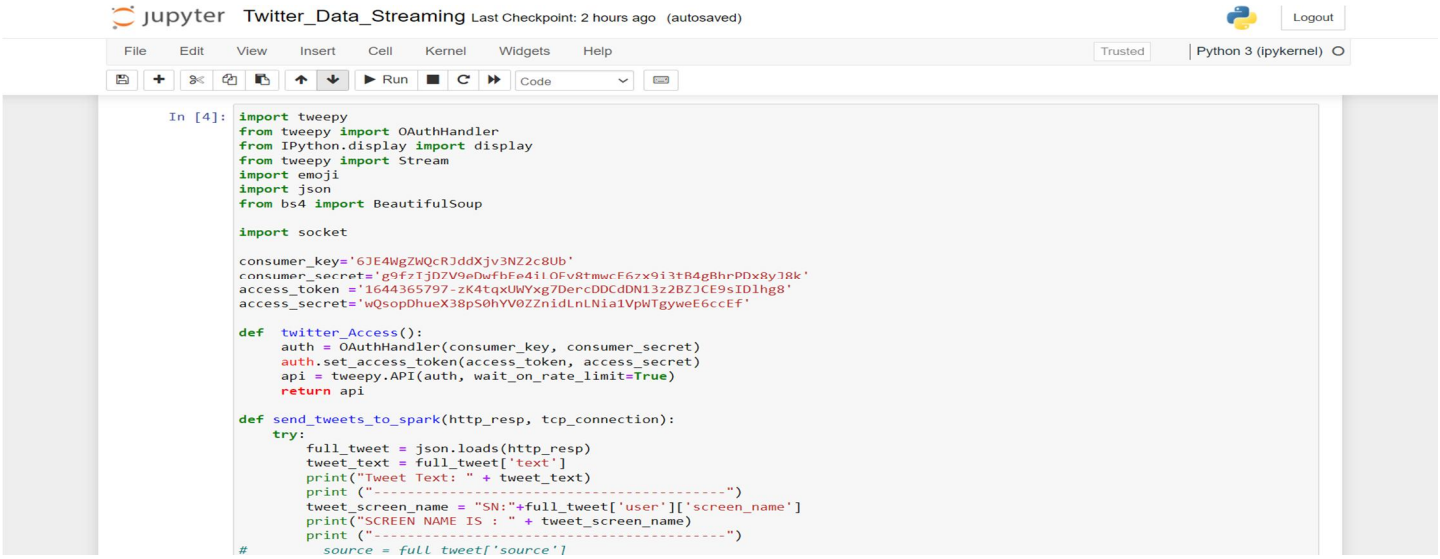
In order to fetch tweets through Twitter API, one needs to register an App through their twitter account.

1) Open the link and click the button: 'Create New App'.
2) Fill the application details. You can leave the callback url field empty.
3) Once the app is created, you will be redirected to the app page.
4) Open the 'Keys and Access Tokens' tab.
5) Copy 'Consumer Key', 'Consumer Secret', 'Access token' and 'Access Token Secret'

© IJRASET: All Rights are Reserved |

B. Building the Twitter HTTP Client

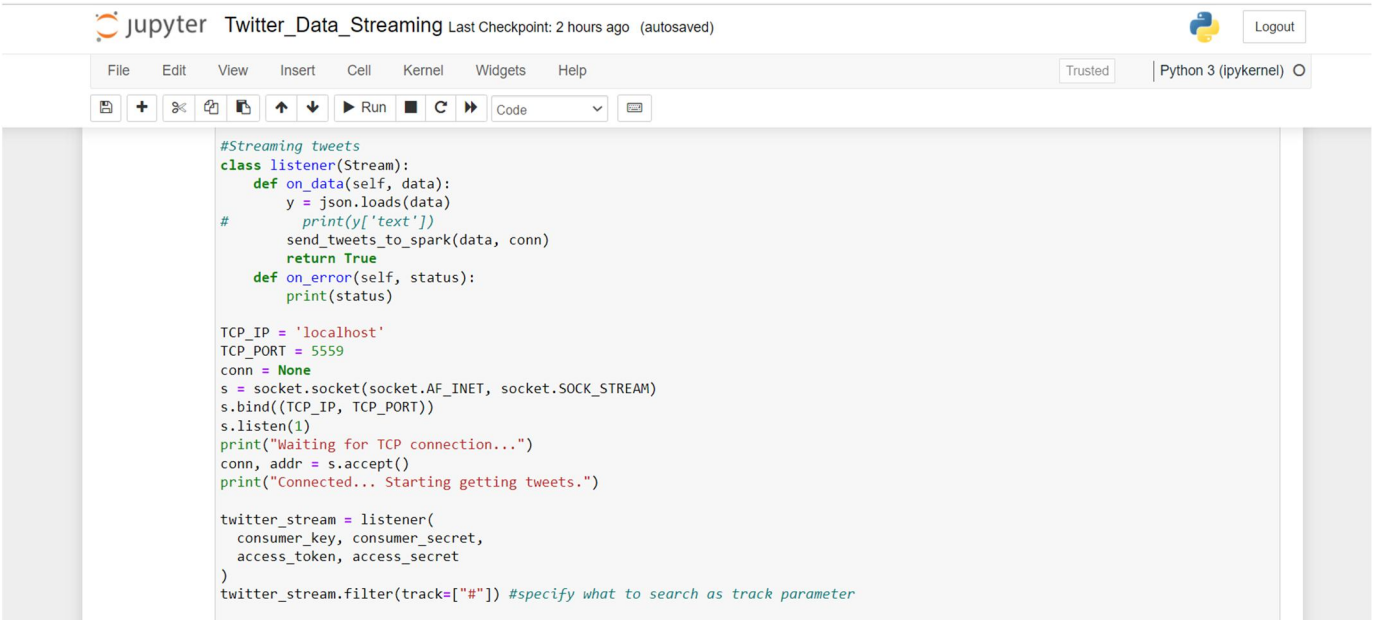
In this step, we get the tweets from Twitter API using Python and pass them to the Spark Streaming instance. First, we created a file called Twitter_Data_Streaming.py and then we'll import libraries i.e import socket, import sys, import requests, import requests_oauthlib,import json and most importantly importing tweepy. Tweepy is publicly released, facilitated on GitHub and enables Python to talk with Twitter stage and utilize its API. Tweepy supports accessing Twitter via Basic Authentication and therefore the newer methodology, OAuth. Twitter has stopped accepted Basic Authentication thus OAuth is currently the sole thing to use the Twitter API



C. Streaming tweets

Streaming the data from the Twitter API requires creating a listening TCP socket in the local machine (server) on a predefined local IP address and port. A socket consists of a server-side, which is our local machine, and a client-side, which is Twitter API. The open socket from the server-side listens for connections from the client. When the client-side is up and running, the socket will receive the data from the Twitter API based on the topic or keywords defined in the Stream Listener instance.

Now, we'll make the app host socket connections that spark will connect with. We configured the IP here to be localhost as all will run on the same machine and the port 5559. Then we'll getting the tweets from Twitter and pass its response along with the socket connection to send_tweets_to_spark for sending the tweets to Spark.

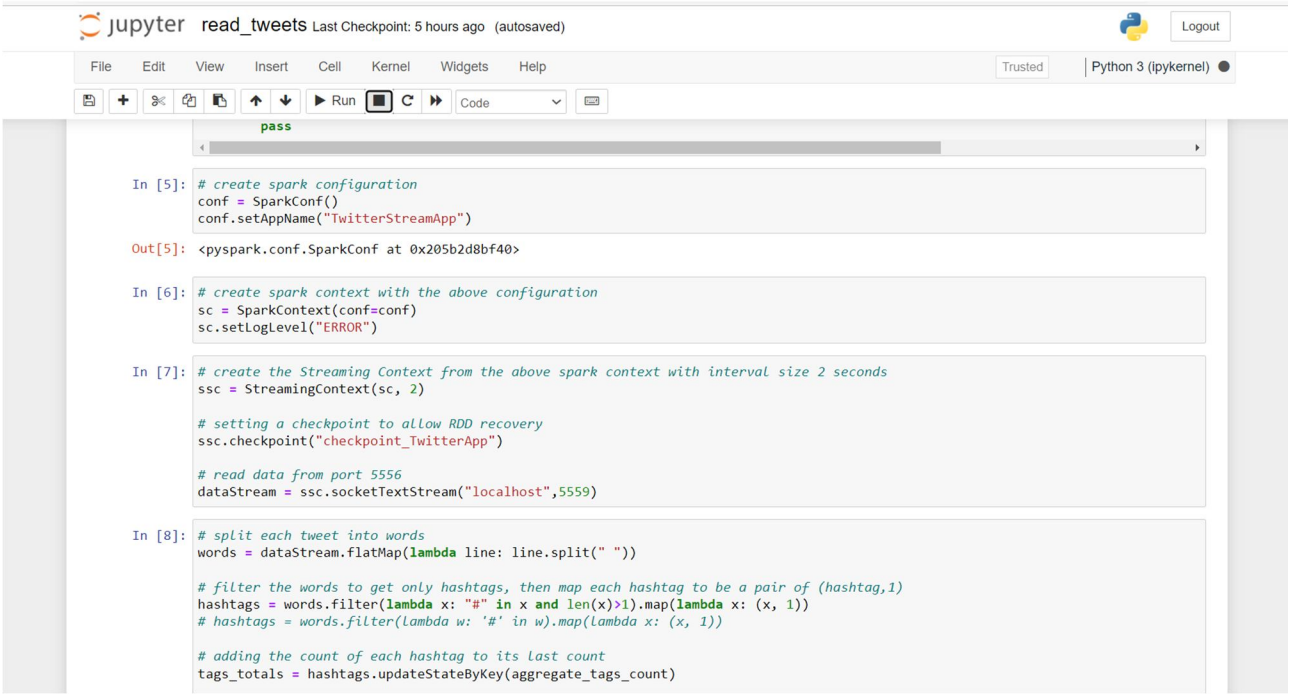


D. Setting Up Our PySpark Streaming Application

In this part we make a different read_tweets.py file which will do real-time processing for the incoming tweets and extract the hashtags from them. Here we import some libraries like SparkConf and Spark Context from PySpark.We created an instance of Spark Context sc, then we created the Streaming Context ssc from sc with a batch interval two seconds that will do the transformation on all streams received every two seconds. We also set the log level to ERROR in order to disable most of the logs that Spark writes.\par

We defined a checkpoint here in order to allow periodic RDD checkpointing .Then we define our main DStream dataStream that will connect to the socket server we created before on port 5559 and read the tweets from that port. Each record in the DStream will be a tweet.We split all the tweets into words and put them in words RDD. Then we filtered only hashtags from all words and mapped them to pair of (hashtag, 1) and put them in hashtags RDD.Then we calculate how many times the hashtag has been mentioned. We need to calculate the counts across all the batches, so we used another function called updateStateByKey, as this function allows us to maintain the state of RDD while updating it with new data. This way is called Stateful Transformation.\par

The updateStateByKey takes a function as a parameter called the update function. It runs on each item in RDD and does the desired logic.In our case, we've created an update function called aggregate_tags_count that will sum all the new_values for each hashtag and add them to the total_sum that is the sum across all the batches and save the data into tags_totals RDD.Then we do processing on tags_totals RDD in every batch in order to convert it to temp table using Spark SQL Context and then perform a select statement in order to retrieve the top ten hashtags with their counts and put them into hashtag_counts_df data frame.Then next step in our Spark application is to send the hashtag_counts_df data frame to the dashboard application. So we'll convert the data frame into two arrays, one for the hashtags and the other for their counts. Then we'll send them to the dashboard application through the REST API.



.

E. Sentiment Analysis of tweets

Here a function is written to firstly fetch the tweets from the API which are related to the word entered by the userand then the function analyzes every tweet. Every word is passed through a function which contains a list of words which are mentioned in a file as bag of words which denote positive ,negative or neutral sentiments. The keyword in the tweet are analyzed from this list of words and then it is decided if its impact is positive ,negative or neutral. Then this returns 0 if neutral,-1 is its negative and +1 if its positive and then it calculates the maximum number of these tweets to conclude if the tweet was positive or negative. The percentage of positive, negative and neutral is also calculated which later on helps to give us the final grade of the tweets of that particular keyword. Lastly a pie chart is displayed which represents the sentiment of the data.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

III. RESULTS

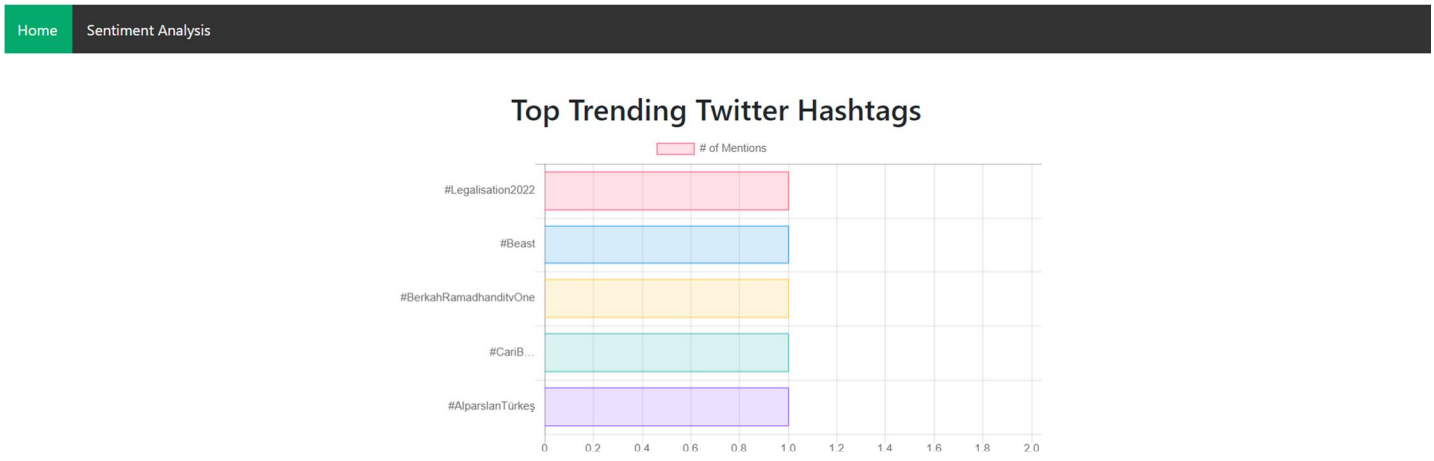


Fig. 3 Front page of the system

The user on opening our webapp is shown a homepage which displays the top trending tweets. Here ,the data analysis of the tweets takes place and it gives us a easy yser friendly ouput in the form of Bar graph. This is a real time analysis which keeps updating itself at constant interval of 2 seconnds and displays the tweet mentions .For better understanding and visual effect we have given different colors to the bar graph representation. When a user hovers over the bar graph he/she can see the number of mentions of that particular hashtag. It keeps updating itself and the highest mentioned hashtag will be at the top most position.

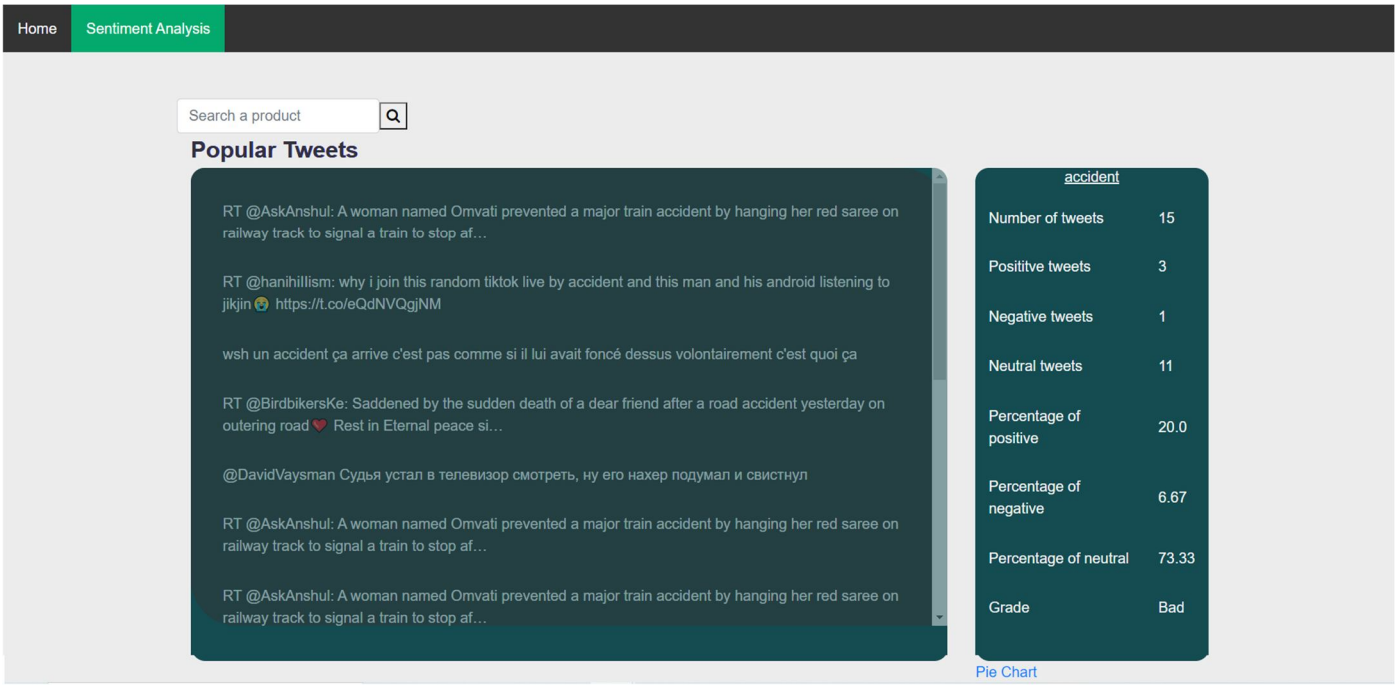


Fig. 4 Twitter Sentiment Analysis of tweets

© IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The second section on our webapp is Sentiment analysis .Once the user clicks on the icon he/she is directed to a new page where any desired word can be given as input in the search box and it will display the results related to that word. All the top recent tweets related to the input word will be shown in the dialog box in the form of tweets. Along with this the sentiment of all the tweets is shown which consists of the total number of tweets displayed, positive tweets, negative tweets and neutral tweets. The percentage of positive, negative and neutral tweets is also displayed for better understanding. Depending on the polarity of the maximum tweets the final sentiment analysis is done and the grade is given for that particular word.

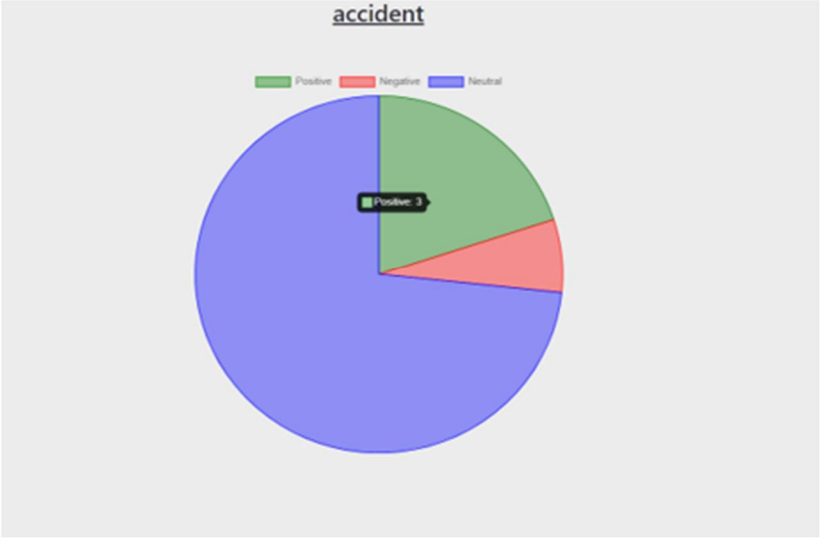


Fig. 5 Pie Chart of Tweets analysed according to sentiments

IV. CONCLUSIONS

Twitter Data in the form of reviews, thoughts, opinion, comments, feedback, and grievance are treated as big data and it cannot be interpreted directly; it should be preprocessed in order to be suitable for mining tasks. In this research, we propose an efficient sentiment analysis technique, utilizing PySpark .The results indicate a significant enhancement in the accuracy of the polarity of the tweets and the additional feature of searching for any desired word is achieved .Along with this we have also did data analyzation on the large Twitter datasets and respresented the real time data of top trending tweets. From the former outcomes, our system can be described as effective and scalable.

V. ACKNOWLEDGMENT

We would like to thank our guide “Prof. Kunal Meher.” who gave us his valuable suggestions and ideas when we were in need of them. He encouraged us to work on this project. We are also grateful to our college for giving us the opportunity to work with them and providing us the necessary resources for the project.

REFERENCES

[1] H. Elzayady, K. M. Badran and G. I. Salama, "Sentiment Analysis on Twitter Data using Apache Spark Framework," 2018 13th International Conference on Computer Engineering and Systems (ICCES), 2018, pp. 171-176, doi: 10.1109/ICCES.2018.8639195.

[2] J. Ranganathan, A. S. Irudayaraj and A. A. Tzacheva, "Action Rules for Sentiment Analysis on Twitter Data Using Spark," 2017 IEEE International Conference on Data Mining Workshops (ICDMW), 2017, pp. 51-60, doi: 10.1109/ICDMW.2017.14.

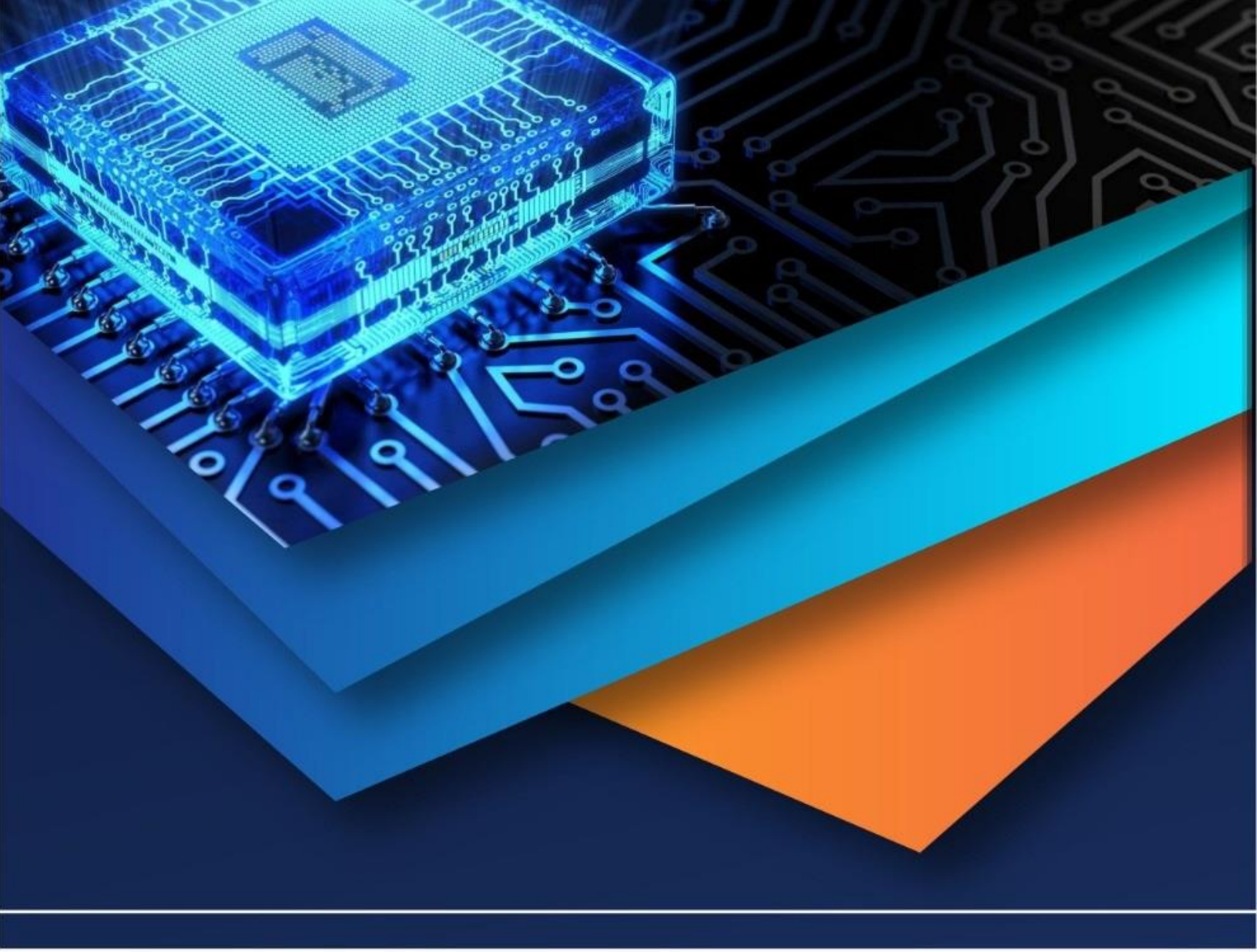
[3] Sentiment Analysis of Twitter Data: A Survey of Techniques(https://www.researchgate.net/publication/324531717_Sentiment_Analysis_Using_Hybrid_Method_of_Support_Vector_Machine_and_Decision_Tree)

[4] R. Plutchick. “Emotions: A general psychoevolutionary theory.” In K.R. Scherer \& P. Ekman (Eds) Approaches to emotion. Hillsdale, NJ; Lawrence Ealrbaum Associates, 1984

[5] P. Basile, V. Basile, M. Nissim, N. Novielli, V. Patti.“Sentiment Analysis of Microblogging Data”. To appear in Encyclopedia of Social Network Analysis and Mining,Springer.

[6] Baltas, Alexandros, Andreas Kanavos, and Athanasios K. Tsakalidis. "An apache spark implementation for sentiment analysis on twitter data." International Workshop of Algorithmic Aspects of Cloud Computing. Springer, Cham, 2016.

© 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)