A NOVEL APPROACH TO ANALYZE SENTIMENT OF SOCIAL MEDIA

Supriya Yerakaraju, [1, a)] Archana Kalidindi, [1, b)]
[1] Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad-500090, Telangana State, India
Email: [a)] supriya.yerakaraju98@gmail.com, [b)] archana.buddaraju@gmail.com

Abstract: Social media is the online platform where everyone can share their opinions. Among the online social media platforms, Twitter is one of the most famous data platforms which holds different kinds of data i.e., structured, semi-structured and unstructured. This raw data can be used for Economic, Industrial, Social, Healthcare, and etc., for processing.
In this paper, we are going to analyze the sentiments of users either positive, negative or neutral. Here we are doing the analysis on basis of the keyword 'Covid'. So, for Covid, how the users are reacting in social media i.e., on Twitter will be analyzed based on the day-to-day life.
The aim of this paper is - we are going to visualize the analysis of sentiments with Positive, Negative and Neutral tweets for Corona Virus. We are going to analyze this using TextBlob.

INTRODUCTION

Now-a-days social media platforms becomes more significant in people's everyday life. Among them, Twitter is one of the most used social media platforms. Twitter is where a user can read the posts and can also write the posts. The twitter messages are also known as Tweets. These tweets will be considered as raw data. We are having so many methodologies to extract those tweets and classify them into Positive, Neutral and Negative sentiments. In these social media platforms, people can discuss and post about their opinions related to various topics, complains about products related to a particular e-commerce platform, current issues in the society and so on. The companies are trying to study their customer's reactions for future improvement of the development of a company by increasing its quality of the products.
For example, a company can be able to know the customer feedback using the sentiment analysis like how they are reacting about a particular product. So that they can further analyze the customer's feedback and their satisfaction based on this sentiment analysis. Hence, Sentiment Analysis became the most famous research area in the computational linguistics.
In this paper, we are using to find and analyze the tweets that are related to 'Covid'. Corona Virus is most spread disease that everyone is aware of. This was first identified in Wuhan, China in December 2019.The disease has spread throughout the world.
Symptoms that lead to covid 19 are cough, headache, fatigue, breathing difficulties, loss of smell and taste. Symptoms may be visible from one to fourteen days after exposure to the virus. People who are infected do not quickly develop noticeable symptoms. And those who develop symptoms are classified as 81% develop mild to moderate symptoms, while 14% develop severe symptoms and 5% suffer critical symptoms. Old age people are at a higher risk of developing severe symptoms. Some people suffer a lot even after months of recovery and damage to organs.
There are many ways to do sentiment analysis and calculating polarity for sentiments. In this paper, we created a Twitter Developer account and collected the following access keys – access key, access token, secret key and secret token. Using these tokens, we are retrieving the tweets from Twitter into Kafka and then processed using Spark. Finally, we are using TextBlob which is a library that returns a Sentiment object. And then we are calculating polarity which gives a value to sentiment from -1.0 (negative) to 1.0 (positive) with 0.0 being neutral.

LITERATURE SURVEY

Several ways have been proposed for Twitter Sentiment analysis from last few years. This part presents a survey that covers Twitter, Kafka, Spark and TextBlob for doing Sentiment analysis. We can also say that, this is an end-to-end data pipeline where we can build a reliable, scalable, and failover mechanism.
This is a continuous pipeline where the Twitter data is coming continuously and Twitter as a producer will produce the data to Kafka continuously until we stop the process or application explicitly or due to some system crash or network issue.
After producing the data to Kafka, Kafka will store the tweets in a topic as JSON objects. These JSON objects will be consumed by Spark using Spark Streaming API and it will do processing and analysis using TextBlob

like calculating polarity score for each tweet. And then using Python library called Matplotlib, we are visualizing the day-wise polarity scores for tweets/sentiments.

There are many cases like making Twitter Sentiment analysis using other different technologies like Flume, Spark, Python, and etc. Among all the other data pipelines, Twitter-Kafka-Spark have the most efficient, scalable, consistency, reliable and failover benefits. We are mainly using this integration to achieve better performance and failover when compared to other technologies.

Social Media based Sentimental Analysis using Hive and Flume

Apache Flume is an open-source, reliable data ingestion tool that is used to collect and aggregate huge volumes of data of unstructured data which can receive data from various sources to HDFS/Hive/Hbase. So here Flume will receive the Twitter data [1] which is an external source and put the data in Flume's channel which is a buffer storage. After that, we do the Sentiment Analysis and then Flume stores the data in HDFS/Hive/Hbase.

The Hadoop Distributed File System

In this paper [4], they used retrieve data using Hadoop component called Flume and stores the tweets in a buffer storage called Channel which is a component of Flume. After that, using Flume there is some preprocessing work is done. Finally, the results are stored in a Data warehouse called Hive which is one of the Hadoop components.

Twitter Sentiment Classification using Distant Supervision

In this paper [5], they did Twitter Sentiment analysis using different Machine Learning classifiers such as Naïve Bayes, Maximum Entropy (MaxEnt), and Support Vector Machines (SVM). But in this paper, they realized that they can use some other Machine Learning or Deep Learning algorithms in order to increase the accuracy of the model.

Twitter Analysis: Twitter Data processing Using Apache Hadoop

In this paper [6], they did social media data analysis like Twitter data analysis using Apache Hadoop. They used Mapreduce framework for propessing and reducing the data. And finally what they analysed is analyzing the tweets that are collected from Twitter and then fetching Tweet IDs of the users whose tweets are collected and found the users who are retweeted most.

Effective Sentiment Analysis on Twitter Data using: Apache Flume and Hive

In this paper [7], they did sentiment analysis by retrieving data from Twitter using Bigdata components named Apache Flume, Hive. Initially, the data is brought from Twitter using Apache Flume which is a data ingestion tool which deals with any kind of data. And then the data is stored in Hive which is a Datawarehouse built on top of Hadoop. Using Hive UDFs and Standford Core NLP, they did Sentiment Analysis.

Bigdata Analysis: Streaming Twitter Data with Apache Hadoop and Visualizing using BigInsights

In this paper [8], they analyzed Twitter data using Hadoop but for streaming Twitter data they used IBM BigInsights. After that they used Mapreduce Framework for analyzing the tweets into Positive and Negative. And then finally, they used bigsheets tool of BigInsights for showing the results with different charts.

Social Media Sentiment Analysis Based On COVID-19

In this paper [9], they analyzed the twitter tweets based on the hashtag related to Covid-19 and to classify the emotions on tweets using Recurrent Neural Network (RNN) and TextBlob. They classified the various texts into not only positive and negative but also weakly positive and weakly negative. Finally, as a result they got emotional classification of the tweets that they received related to particular topics.

PROPOSED SYSTEM

In this paper, we are developing an application that parses, preprocesses, processes and analyzes the tweets and calculates polarity score for each tweet/sentiment by dividing them as Positive, Negative and Neutral.

Initially, we created a Twitter Developer account and saves the access tokens. Creating a Kafka Producer application that produces tweets into Kafka topic. After that, those tweets were parsed and preprocessed and then stored as JSON objects in that Kafka topic. And then Spark as a consumer, the JSON objects were consumed by a Spark application that is processing the tweets and stores those processed tweets in Spark dataframe.

And then finally, by using a library named TextBlob, we are analyzing the tweets as sentiments and giving them a polarity score for each sentiment. The polarity scores are given as -1.0 for negative sentiment, 1.0 for positive sentiment and 0.0 for neutral.

And then finally, we visualized the results how these polarity scores were either increasing or decreasing daily in a graph using Python matplotlib library.

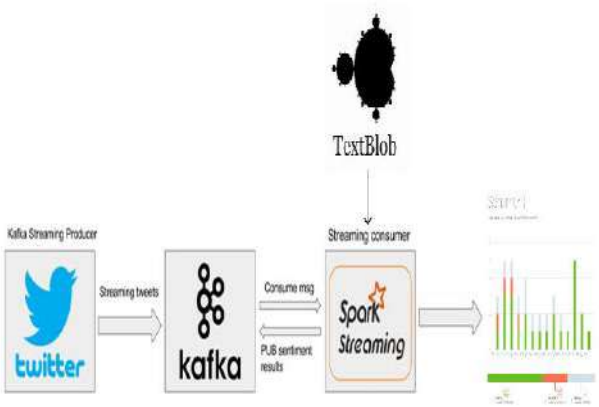

FIGURE 1: *Architecture Diagram for Sentiment Analysis of social media*

Initially, we are retrieving data from Twitter using Twitter Stream Listener API and by using the access tokens that are provided by Twitter when creating Twitter Developer account, the tweets will be retrieved based on a particular hashtag into Kafka topic. And then the tweets will be parsed and preprocessed using Kafka APIs. After that Spark as a consumer consumes the tweets from Kafka and then it will do some processing and stores the results in a Spark Dataframe. After that using TextBlob library, we are analyzing the tweets as sentiments and calculating the polarity score for Positive, Negative and Neutral sentiments. Finally, using Python visualization library named Matplotlib, we are visualizing the results in a graph.

IMPLEMENTATION

Any paper can be analyzed based on a dataset. For our thesis, we have named it as Sentiment Analysis of Social Media.

For Sentiment Analysis of Social Media, we used real-time data from Twitter which produces continuous tweets until we explicitly stop the data receiving.

To implement this paper, we used the technologies like Kafka, Spark with Python API, TextBlob and Matplotlib. In addition to that, we used the tools and softwares like Oracle Vitual Box, Apache Hadoop VDI, TigerVNC viewer, Kafka with 2.11 version, Spark 2.3.4 with Hadoop2.7 version, Python3.7, Java1.8, and Anaconda Navigator.

Initially, we downloaded and installed the Oracle Virtual Box, TigerVNC viewer, Anaconda Navigator for Linux OS and Apache Hadoop VDI. After that we created an Apache Hadoop machine with 10GB, 2 CPU cores with network option – Bridge adapter for receiving tweets from Twitter i.e., live data/real-time data and started all the Hadoop daemons.

After that, we used Kafka which is a distributed messaging system for bringing real-time data from different sources into different destinations. In our paper, our source is Twitter i.e., live streaming Twitter data and destination is Spark Dataframe. We stored the received tweets in a Kafka topic. For that purpose, we wrote a Kafka producer application that receives data from Twitter by creating a Twitter Developer account and using the access tokens in that Kafka producer application, pushing them into a Kafka topic based on the hashtag. In this paper, we are retrieving the tweets related with '#Covid' hashtag. So that application will retrieve Covid related tweets into the Kafka topic.

After that, we parsed and preprocessed the Twitter tweets to do the Sentiment analysis. We're using Spark which is a unified computing engine used for processing the data. We used Spark as a Kafka Consumer which will receive the parsed Twitter data from Kafka topic and do the processing.

Now after processing the data, we used TextBlob which is built on top of NLTK which offers a lot of features such as Sentiment Analysis, Noun-Phrase extraction, postagging and etc.

Using the TextBlob, we classified the parsed and processed Twitter tweets into sentiments like positive or negative or neutral. This library will classify tweets into sentiments.

After classifying the tweets into sentiments using TextBlob, we are calculating the polarity scores for those sentiments with values 1.0 for positive, -1.0 for negative and 0 for neutral.

After that, we plotted a pie chart based on polarity scores of positive, negative and neutral sentiments in Sentiment Analysis of Social Media using BigData.
Likewise, we did Sentiment Analysis of Social Media using RNN and plotted a pie charted based on the polarity scores of positive, negative and neutral.
And we compared the performance of Sentiment Analysis of Social Media using Bigdata as well as using RNN. We observed the better performance and better quality using BigData.
Finally, we are using a Python library called Matplotlib for visualizing the how these polarity scores are increasing and decreasing. The resultant graph will look like this –

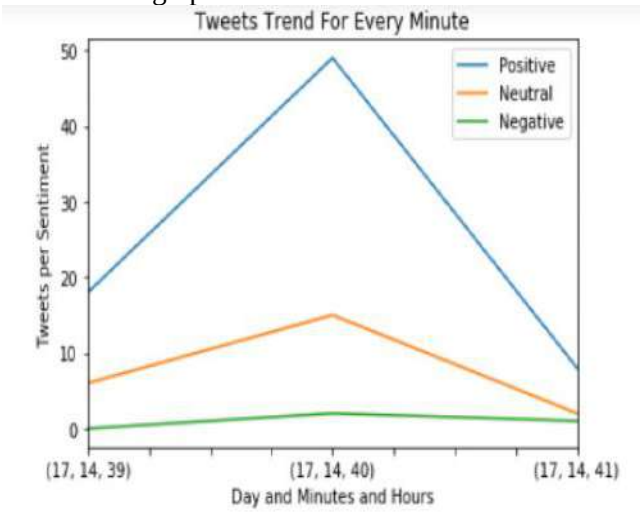

FIGURE 2: *Architecture Diagram for Sentiment Analysis of social media*

EXPERIMENTATION AND RESULTS

In our paper, we have used the Bigdata technologies like Spark and Kafka for dealing with huge amounts of data as well as the live streaming data from Twitter.
Basically, there are so many papers on Sentiment Analysis using different technologies but no one did using Bigdata technologies that we have used.
We used these technologies, because Kafka and Spark will have features like fault-tolerant, reliable, fast retrieval of live data.
The other thing that we experimented is we used TextBlob which NLTK based library for calculating the polarity score.
We had done Sentiment Analysis using the RNN alogorithm also but we noticed that using Bigdata technologies and TextBlob has given the better result when compared to RNN.
The resultant graph after calculating polarity score for each sentiment using TextBlob is:

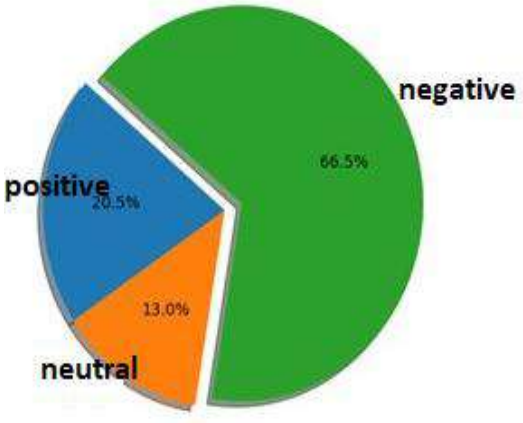

FIGURE 3: *Sentiment Analysis of social media polarity graph using Big data – TextBlob*

The resultant graph after calculating polarity score for each sentiment using RNN is:

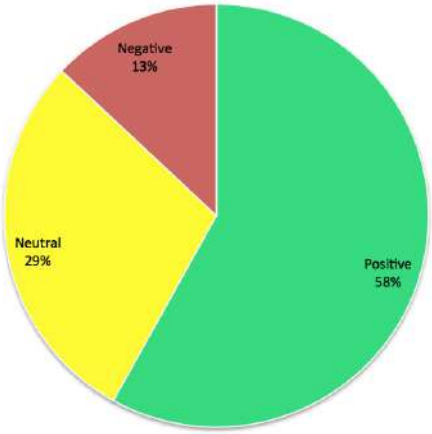

FIGURE 4: *Sentiment Analysis of social media polarity graph using RNN*

The comparison graph for polarity scores of Positive, Negative and Neutral sentiments based on polarity scores using RNN and BigData pipeline with TextBlob is as shown in the below figure –

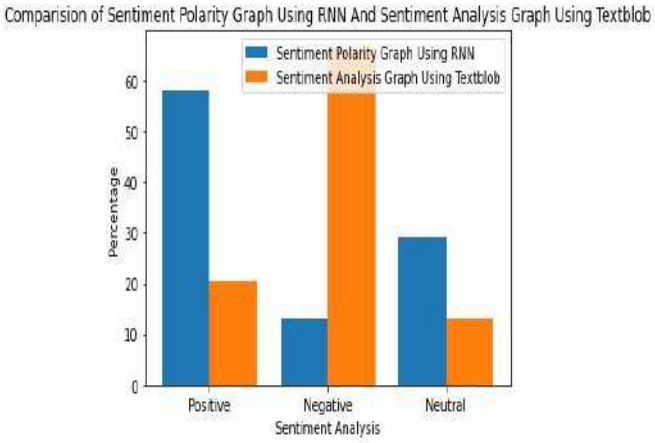

FIGURE 5: *Comparison graph for Sentiment Analysis of social media using Big data - TextBlob and RNN*

CONCLUSION AND FUTURE ENHANCEMENTS

In this paper, we retrieved, parsed, preprocessed and processed the Twitter tweets by producing the tweets into Kafka using access tokens. We filtered the tweets based on hashtag (#covid). And then Spark consumed those tweets that are stored as JSON objects using Spark Streaming API. After consuming those tweets, we analyzed the tweets and calculated polarity using TextBlob library and visualized the tweets as Positive, Negative and Neutral using Matplotlib library.

This approach is basically having a sample sentiment analysis. So, even though we are using continuous real-time streaming data, we can do more analysis like emotion analysis. Basically, when compared to other Big-data technologies, using Kafka and Spark is the efficient way for data analysis.

Basically people do the sentiment analysis using traditional research processes which gives the polarity result very good but using Bigdata technologies will fast and can deal with large amounts of data.

And also the people are simply taking datasets of Twitter data and doing the analysis but in this paper, we are dealing with real time streaming data from Twitter (All the latest tweets related to that hashtag will be retrived and stored).

In future, we can do more analysis part using this data pipeline such as Emotional analysis. There we can use Spark ML library for analyzing the tweets based on Positive, Negative and Neutral as we are using Textblob library. Instead of Twitter tweets, we can use the images with different emotions like fear, angry, happy, sad, and etc. And also finally we can store those results in a Data warehouse called Hive.

REFERENCES

1. Pooja S. Patil1, Pranali B. sable, *Sentiment Analysis on Twitter Data Using Apache Flume and Hive*, IRJET, Feb-2016.

2. Jianqiang, Z., Xiaolin, G., & Xuejun, Z. (2018). *Deep convolution neural networks for twitter sentiment analysis*. IEEE Access, 6, 23253–23260https://doi.org/10.1109/ACCESS.2017.2776930
3. Sunil B. Mane, Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde, *Real Time Sentiment Analysis of Twitter Data Using Hadoop*, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3098 – 3100.
4. K. Shvachko, H. Kuang, S. Radia, and R. Chansler, *The Hadoop Distributed File System*, in the 26th IEEE Symposium on Mass Storage Systems and Technologies, pp. 1-10, May 2010.
5. Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision*. CS224N Project Report, Stanford, 1-12.
6. International Journal Of Core Engineering & Management (IJCEM), *Tweet Analysis: Twitter Data processing Using Apache Hadoop*, Volume 1, Issue 11, February 2015.
7. Penchalaiah.C, Murali.G Suresh Babu.A, *Effective Sentiment Analysis on Twitter Data using: Apache Flume and Hive*, Computer Science and Engineering Dept, JNTUACEP, Pulivendula, Vol. 1 Issue 8, October 2014.
8. Manoj Kumar Danthala, Dr. Siddhartha Ghosh, 2015, *Bigdata Analysis: Streaming Twitter Data with Apache Hadoop and Visualizing using BigInsights*, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 04, Issue 05 (May 2015), http://dx.doi.org/10.17577/IJERTV4IS050643
9. Laszlo Nemes & Attila Kiss (2021) *Social media sentiment analysis based on COVID-19*, Journal of Information and Telecommunication, 5:1, 1-15, DOI: 10.1080/24751839.2020.1790793.
10. Bahrainian, S.A., Dengel, A., *Analysis of Sentiment using Sentiment Features*, In the proceedings of WPRSM Workshop and the Proceedings of IEEE/WIC/ACM International Conference on Web ntelligence, Atlanta, USA, 2013.
11. Manoj Kumar Danthala, *Tweet Analysis: Twitter Data processing Using Apache Hadoop*, International Journal of Core Engineering & Management (IJCEM).
12. Xu, J., Huang, F., Zhang, X., Wang, S., Li, C., Li, Z., & He, Y. (2019). *Sentiment analysis of social images via hierarchical deep fusion of content and links*. Applied Soft Computing, 80, 387–399. https://doi.org/10.1016/j.asoc.2019.04.010
13. Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., & By, T. (2012). *Sentiment analysis on social media*. In 2012 IEEE/ACM international conference on advances in social networks analysis and mining(pp. 919–926).
14. Ortis, A., Farinella, G. M., Torrisi, G., & Battiato, S. (2018). *Visual sentiment analysis based on on objective text description of images*. In 2018 international conference on content-based multimedia indexing (CBMI) (pp. 1–6).
15. Arras, L., Montavon, G., Müller, K. R., & Samek, W.2017). *Explaining recurrent neural network predictions in sentiment analysis*. Preprint arXiv:1706.07206.
16. Balahur, A. (2013). *Sentiment analysis in social media texts*. In Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis (pp. 120–128).
17. Sentimental Analysis, Inc. [Online]. Available: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis [Accessed 23 March 2013].
18. Leskovec, J. (2011). *Social media analytics: Tracking, modeling and predicting the flow of information through networks*. In Proceedings of the 20th international conference companion on world wide web (pp. 277–278).
19. Pandey, A. C., Rajpoot, D. S., & Saraswat, M. (2017). *Twitter sentiment analysis using hybrid cuckoo search method*. Information Processing & Management 53(4),764–779. https://doi.org/10.1016/j.ipm.2017.02.004 [Crossref], [Web of Science ®], [Google Scholar].
20. Wang, Y., & Li, B. (2015). *Sentiment analysis for social media images*. In 2015 IEEE international conference on data mining workshop (ICDMW) (pp. 1584–1591). [Crossref], [Google Scholar].

