**Sentiment Analysis using Distributed Machine Learning**

CSC-721 Distributed Systems

Alisha Karna

*Department of Computer Science*

*University of South Dakota*

Vermillion, South Dakota

alisha.karna@coyotes.usd.edu

Madhavi Awasthi

*Department of Computer Science*

*University of South Dakota*

Vermillion, South Dakota

madhavi.awasthi@coyotes.usd.edu

Srijana Raut

*Department of Computer Science*

*University of South Dakota*

Vermillion, South Dakota

srijana.raut@coyotes.usd.edu

**Abstract:**

The rapid growth of social networking and marketing sites in the Internet era has become a major concern for most businesses. In today's world all businesses have their online portal through which they are monitoring their customer needs so that they can generate good insights and improve their customer service for better business model plans. Thus, Sentiment analysis tools are essential to detect and understand customer feelings. Thus, we find the problem as understanding those concepts and dealing with them is one of the major tasks. In our project, we are going to perform text mining where the data can be from various sources. Our data is in semi structured format i.e which needs to convert it into structured format by performing some ETL (Extract, Transform and Load) techniques. For ETL and ML building, we will be using Distributed Framework Apache Spark. Finally, we deploy our streamlit web app in AWS EC2 Instance. We can visualize the emotion from the respective text/sentence with the prediction and confidence score and these are stored in a database Deta.

*Keywords: Apache Spark, HDFS, AWS EC2 instance, AWS S3 bucket, Deta Database, Streamlit*

**Introduction:**

Nowadays, social networking services turn the entire world into a small village where users may share their opinions, sentiments, experiences, and advice for others to benefit from. Since so many people use social media regularly, there are a ton of comments, opinions, and articles out there. [1] Finding an automated method for analyzing and categorizing user opinions on social networks could be very important. This is mostly due to the fact that it is regarded as an excellent tool for obtaining direct feedback or information from people. Sentiment Analysis is the process of categorizing texts or documents according to their polarity. It is used to analyze raw text to drive objective quantitative results using natural language processing, machine learning, and other data analytics techniques and detect different sentiment in text, and often businesses use it to gauge branded reputation among their customers.

Social media currently has a large variety of text posts, and this number keeps growing daily. As a result of this tremendous data expansion, the existing databases are unable to manage a large volume of data quickly. Additionally, although these databases are made to handle structured data, they have limitations when working with large amounts of data. As a result, enterprises find it difficult to handle and process vast amounts of unstructured data using conventional technologies. In this project, we are going to analyze the sentiment of the text in which the data can be from various sources. [2] Our data is in semi structured format i.e which needs to convert it into structured format by performing some ETL (Extract, Transform and Load) techniques. For ETL and ML building, we will be using Distributed Framework Pyspark. One of the most effective manners is the parallel computing techniques when dealing with big data, which include distributed computing. After Preprocessing the data is then stored on a Hadoop distributed file system. The clean data is then extracted from HDFS and we build the Machine learning model. For visualization, we use streamlit and deploy on AWS. Finally, the predicted data is stored in the database Deta.

**Motivation:**

Before explaining the motivation to start the project, let’s dive into a few problems that we addressed using Distributed Machine Learning techniques. Large amount of data needs to be processed, stored and cleaned. Likewise, large amount of time is required for a machine to process the whole data. Maintaining a centralized server is difficult. Data scalability, security and reliability is also a major problem. For all these problems, we have come up with an idea of Distributed Machine Learning for Sentiment Analysis.

First of all, let’s talk about Apache Spark. Apache Spark is the distributed processing framework which we follow for our data preprocessing and Model building. It also processes the larger dataset and is known as RDD i.e Resilient Distributed Dataset. RDD is the building block of Spark application which is fault tolerant in nature and is 100 times faster than pandas.

Here, in our project, we have analyzed the sentiment of the twitter text where we used Amazon s3 bucket and Hadoop Distributed File system for storing the data. And Pyspark as Distributed Framework for model building which is then deployed on AWS EC2 Instance.

**Related work and Literature Review:**

We had to conduct extensive research and study to complete this project as it makes use of the PySpark library and applies the principles of distributed systems. We discovered certain materials based on the implementation of the PySpark library with respect to the concepts of distributed systems into actual applications after going through a large number of research papers, web publications, and journals. Here are a few works that are similar to our project and significantly aid in its completion.

**Sentiment Analysis on Twitter Data using Apache Spark:**

Both the research and commercial worlds are finding sentiment analysis to be an intriguing topic. Sentiment describes a person's thoughts or sentiments regarding a particular problem. [3]Additionally, it's regarded as a direct application for opinion mining. Twitter is a rich source of textual data and one of the most important data volumes thanks to the enormous volume of tweets that are written down every day; as a result, this data has a variety of goals depending on the amount of data needed and the processing that will be required, such as business, industrial, or social goals. [4]Big data, which refers to the vast amount of data that is growing at a rapid rate every second, requires the use of specialized processing methods and powerful computers to carry out the necessary mining operations. In this study, they have used the Apache Spark framework, which is regarded as an open-source platform for distributed data processing and made use of distributed memory abstraction, to execute a sentiment analysis. [5]To successfully handle a massive amount of data, Apache Spark's Machine learning library (MLIB) is used. In order to improve the accuracy of sentiment analysis classification, they took a few preprocessing and machine learning text feature extraction stages. When using the Naive Bayes, Logistic Regression, and Decision Tree classification algorithms, their proposed approach outperforms existing approaches in terms of classification outcomes. Last but not least, their solution calculates Apache Spark's scalability performance.

**Sentiment Analysis of Online Food Reviews using Big Data Analytics:**

Sentiment analysis has grown significantly in importance in recent years, being mostly applied to large datasets and helping researchers in the application of methodologies. They used Big Data to solve this issue because traditional solutions are unable to handle the exponential growth of Amazon's food data. [6] In this study, they investigated various approaches to sentiment analysis for large datasets of Amazon Fine Food reviews using the Apache Spark data processing engine. By utilizing MLlib, the machine learning package for Apache Spark, three techniques-Linear SVC, Logistic Regression, and Naive Bayes-that have greater than 80% accuracy are implemented. When these techniques are used, they found that linear SVC outperforms NB and logistic regression in terms of performance.

**Proposed Methodology:**

1. **Data Collection**:

In this project, we are going to analyze text mining where the data can be from various sources. The dataset has been taken from the kaggle. There are various texts in the datasets such as happy, anger, sad, etc. The text including the terms "happy," "joy," etc. can be categorized as positive remarks whereas, the text containing the words "anger," "sad," etc. can be categorized as negative comments in the given dataset.

**2.   Tools Used:**

The different tools employed in the proposed methodology is described below:

1. **PySpark:**

PySpark is a Python API for Spark released by the Apache Spark community to support Python with Spark. Pyspark is considered as the distributed Framework which relies on ApacheSpark. ApacheSpark is built with a Resilient Distributed System. PySpark provides us with more and better solutions to deal with big data and also gives us flexible and scalable tools to exploit big data and gain its benefits.

1. **AWS Bucket:**

An Amazon S3 bucket is a public cloud storage resource in AWS S3, an object storage device.  The objects that are stored in Amazon S3 buckets, which resemble file folders, are made up of data and the metadata that describes it.

1. **HDFS:**

HDFS stands for Hadoop Distributed Files System. It is a distributed file system that handles large data sets running on commodity hardware and used to scale a single Apache Hadoop cluster to hundreds (and even thousands) of nodes.

**d. Amazon EC2:**

Amazon EC2 (Amazon Elastic Compute Cloud) is a web service that offers secure, scalable compute capability in the cloud. It is intended to make web-scale cloud computing simpler for the developers.

1. **Streamlit:**

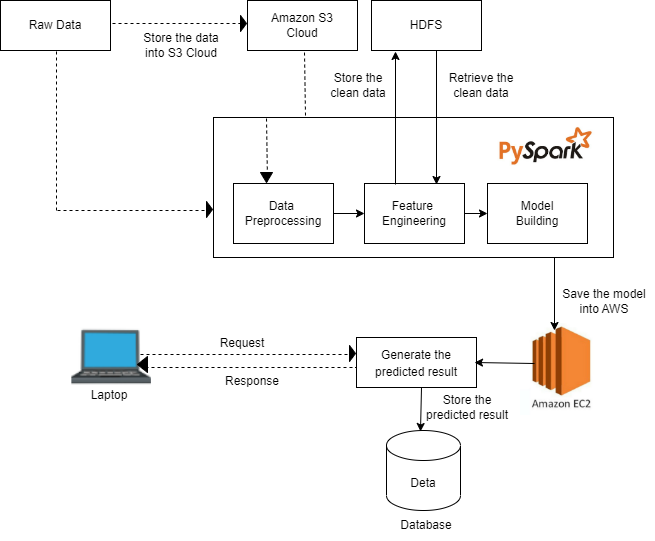
Streamlit is an open-source app framework in Python that helps us to create web-apps for data science and machine learning in a short period of time.

1. **Deta:**

Deta is a cloud-based Database where we can store the data and easily perform the CRUD operations. It is free and easy to implement with python, Nodejs. We use Deta for storing the predicted score with the original text.

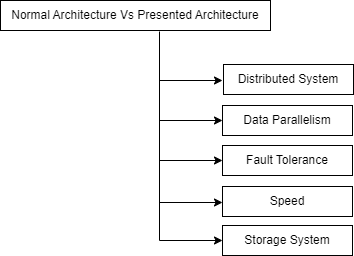
**Implementation:**

At first, the raw data are stored at Amazon S3 Cloud and are extracted to preprocess using Distributed Framework called Pyspark which is the Python API for Apache Spark. Pyspark is 100 times faster and thus the clean data are stored on Hadoop Distributed File System as a chunk. We use HDFS because of its easy integration with Distributed Framework called Pyspark. After then data is loaded from HDFS and trained model is visualized using streamlit. The streamlit web app is then deployed on AWS EC2 instance server. Finally, the query is hitted in from the text and final predicted score are stored on Cloud based NoSql database called Deta where we can also perform all the CRUD operations.

****

*Fig 1: Proposed Architecture*

**Comparison of Proposed Architecture with Normal Architecture:**

****

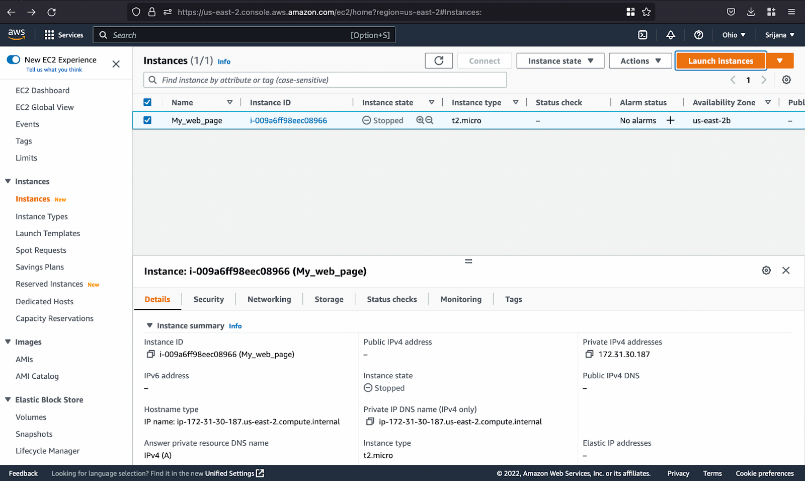
*Fig 2: Comparison of Proposed Architecture with Normal Architecture:*

* Distributed system performs the computations

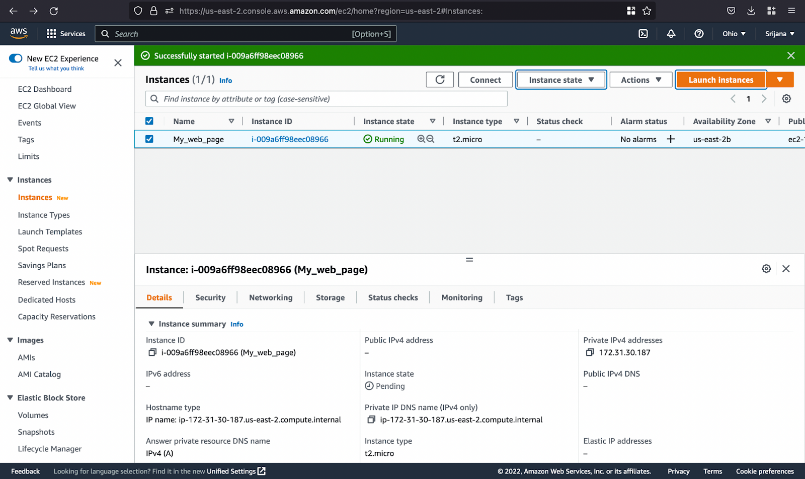
in distributed manner whereas, in normal architecture it does not provide computation in a distributed manner.

* For the storage of the data, we have used Amazon S3 cloud whereas SFTP server is used in normal architecture.
* Now talking about the speed, proposed architecture is very fast because large tasks are split across the different nodes and are scheduled into the cluster whereas, normal architecture is slow because tasks are not splitted.
* In our architecture, data parallelism is achieved by processing the data in parallel with multiple workers which is not possible in normal architecture.
* For the fault tolerance, we have used Apache Spark which is fault tolerant in nature and can help to recover the loss after the failure occurs which is not possible in normal architecture because there is single edge deployment resulting in single point of failure.

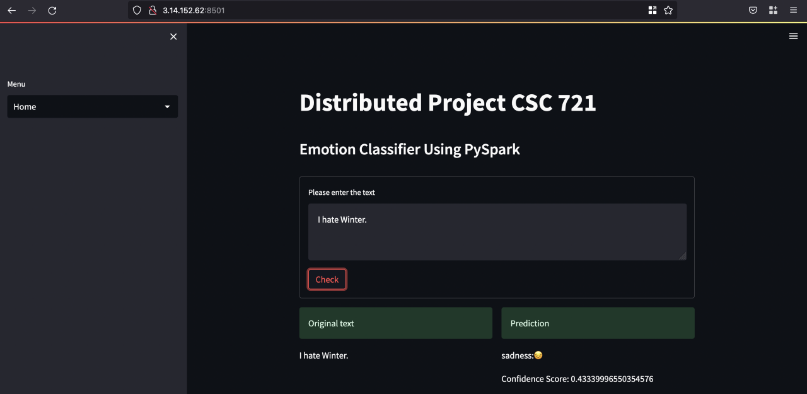
**Results:**

****

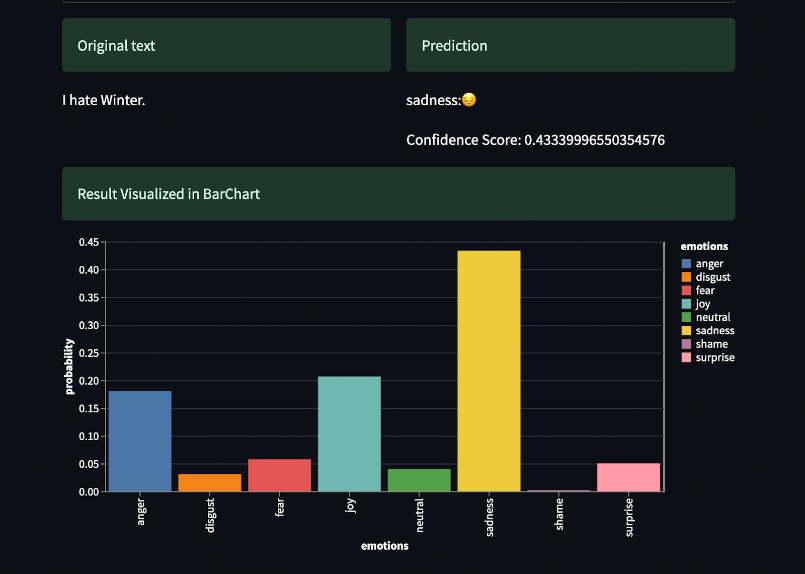
*Fig 3: AWS Instance Stopped*

****

*Fig 4: AWS Instance Running*

****

*Fig 5: Query an emotion prediction on streamlit dashboard*

****

*Fig 6: Result with visualization graph on streamlit dashboard*

**Conclusion:**

A huge amount of data has been generated on social media platforms, which contains crucial information for various applications. As a result, sentiment analysis is critical for analyzing public perceptions of any product or service. Thus, Sentiment analysis tools are essential to detect and understand customer feelings. There are lots of experiments done in the domain of twitter data such as abusive data or others only few are done in the process of   sentiment analysis. In contrast, we proposed a multi-class sentiment analysis dataset using the distributed Pyspark framework.

**References:**

|  |  |  |
| --- | --- | --- |
| **[1]** |  | **A. A. B. X. I. V. O. R. R. Passonneau, "Sentiment Analysis of Twitter Data," in *Proceedings of the 12th Conference of the European Chapter*, New York, NY 10027 USA, 2009.** |

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **[2]** |  | **S. W. F. W. J. L. Y. H. Z Yuan, "Domain attention model for multi-domain sentiment classification," National Key Research and Development Program, China, September 2018.** |
| **[3]** |  | **K. M. B. I. S. Hossam Elzayady, "Sentiment Analysis on Twitter Data using Apache Spark Framework," 2018 13th International Conference on Computer Engineering and Systems (ICCES), MTC, Cairo, Egypt, December 2018.** |
| **[4]** |  | **A. C. S. R. &. C. K. Mayur Wankhade, "A survey on sentiment analysis methods, applications, and challenges," Springer Nature, Bangalore, 560078, India, February 2022.** |
| **[5]** |  | **P. Perrin, "Inferring User Emotions in Texts Using SparkNLP," Holler developers, United state, Apr 22, 2021.** |
| **[6]** |  | **M. J. A. N. S. K. A. Y. H. M. F. S. Hafiz Muhammad Ahmed, "Sentiment Analysis of Online Food Reviews using Big Data Analytics," Elementary Education Online, Lahore, Pakistan , Apr 2021.** |