**Project Proposal: Big Data Analytics on Social Media Usage and Emotional Well-being**

**Project Title: Analysing the Impact of Social Media Usage on Emotional Well-being Using Big Data Technologies**

**Prepared By:**

* **Bratati Chakraborti**

**Guided By:**

* **Prof. Axel Wemmel**

**Objective:**

The primary objective of this project is to perform a comprehensive data analysis on using a range of big data technologies on a local machine.

The primary objective of this project is to analyze the usage patterns of different social media platforms (Facebook, LinkedIn, Snapchat, Twitter, etc.) by males and females and examine their effects on various aspects of mental health, including happiness, boredom, sadness, and anxiety. The project will utilize a range of big data technologies on a local machine to process, clean, join, analyze, and visualize the data to derive meaningful insights.

Data source: multiple CSV files from Kaggle ([Social Media Usage and Emotional Well-Being (kaggle.com)](https://www.kaggle.com/datasets/emirhanai/social-media-usage-and-emotional-well-being/) )

**Tools and Technologies:**

1. **Hadoop**:
   * **HDFS (Hadoop Distributed File System)**: For storing data on a local machine across a simulated cluster.
   * **MapReduce**: Used for processing large data sets on a local machine through WSL (Windows Subsystem for Linux) with Ubuntu.
2. **Apache Spark**:
   * **PySpark**: Python API for Spark to perform data processing and analysis in Jupyter Lab.
   * **Spark SQL**: For structured data processing and running SQL queries.
   * **MLlib**: Spark's machine learning library for implementing regression analysis.
3. **Pandas**:
   * A powerful Python data analysis toolkit used for cleaning and transforming the data.
4. **Jupyter Lab**:
   * An interactive development environment for data analysis and visualization.
   * Utilized for writing and running Spark code, and creating notebooks that combine live code, equations, visualizations, and narrative text.
5. **Visualization Tools**:
   * **Matplotlib**, **Seaborn**, and other libraries for creating static, animated, and interactive visualizations, including tree maps, chord charts, and 3D animations.

**Project Description:**

1. **Data Ingestion and Storage**:
   * The project begins with the ingestion of training and testing CSV files from Kaggle into the Hadoop ecosystem using HDFS on a local machine.
   * Additional CSV files, such as subscription data, are also ingested and stored in HDFS for efficient retrieval and further processing.
2. **Data Processing with Hadoop MapReduce**:
   * Using the MapReduce paradigm on the local machine via WSL with Ubuntu, the project processes the test data CSV to perform initial cleaning, transformation, and aggregation tasks.
   * The resultant CSV from the MapReduce process is stored back in HDFS for further analysis.
3. **Data Cleaning and Integration with Pandas**:
   * Initial data cleaning and transformation are performed using Pandas on the training data. This includes handling missing values, filtering, and creating a processed CSV.
   * Different join operations (inner, self, left outer, right outer, full outer join) with subscription CSV files are performed to integrate various data sources using Spark SQL.
4. **In-Memory Data Processing with Spark**:
   * Spark is used to load the processed and integrated data from HDFS.
   * Advanced data manipulations and aggregations are carried out using Spark SQL.
   * The MLlib library is utilized to apply regression analysis and limited scope pattern matching to identify trends and patterns in the data.
5. **Visualization and Analysis with Jupyter Lab**:
   * The results from the Spark computations are imported into Jupyter Lab.
   * Visualization libraries like Matplotlib and Seaborn are used to create detailed and interactive plots, including tree maps, chord charts, and 3D animations, to represent the data insights.
   * These visualizations will focus on illustrating how the usage of different social media platforms by males and females correlates with various mental health outcomes such as happiness, boredom, sadness, and anxiety.
6. **Future Scope**:
   * Explore the integration of more advanced machine learning algorithms and AI techniques to further enhance the predictive capabilities and insights derived from the data.

**Expected Outcomes:**

* Efficient processing and storage of large data sets using HDFS on a local machine.
* Enhanced data processing speed and performance using Hadoop MapReduce and Apache Spark.
* Detailed and interactive visualizations providing clear insights into the data.
* Insights into how different social media platforms impact mental health differently for males and females.
* Foundation for future integration of advanced machine learning and AI techniques.