Advanced Fake News Detection System by using PySpark

**Abstract:**

The Advanced Fake News Detection System is a comprehensive project which is aimed at combating the day by day increasing issue of fake news and misinformation in today's digital media. We hope to implement this project by making use of PySpark. The idea behind the project is simple, with growing dependency on digital media, the need to improve information integrity and avoid the spread of fake news is necessary and thus, this project leverages current leading Natural Language Processing and Machine Learning techniques for classifying news posts and social media posts as 'real' or 'fake'. The project involves making use of publicly available datasets, ie Kaggle's Fake and Real News Dataset. The action plan is to complete the project under 6 weeks, starting by collecting the data from publicly available datasets like mentioned as Kaggle's Fake and Real News Dataset. Then after the cleanup or preprocessing of data, we will be extracting features using TF-IDF. Post this we will build and train machine learning models in order to detect the difference between a real and a fake news. Lastly, the system will be evaluated on the basis of metrics such as accuracy, precision, recall, F1-score, ROC-AUC and maybe more metrics which we might encounter as we dive deep into it. With the integration of NLP and ML, along with the evaluation metrics to make sure the model runs smooth and efficiently, this project will offer a reliable and scalable solution for identifying fake news to serve the greater cause of establishing trust in digital information.