Fake profile recognition using big data analytics in social media platforms

Name: Suvam Das Roll No: 22N0454

Abstract

Spark ML (Machine Learning) is a component of Apache Spark, which is an open-source big data processing framework. Spark ML provides a set of high-level APIs and tools for building scalable and distributed machine learning pipelines on top of Spark. Spark ML supports a wide range of machine learning algorithms and tools, including: Data preparation and transformation: Spark ML provides a set of transformers and feature extractors for processing and transforming data, such as Vector Assembler for combining multiple features into a single vector, String Indexer for converting categorical variables into numerical indices, and Min Max Scaler for scaling features to a given range. Model training and tuning: Spark ML provides a set of algorithms for training supervised and unsupervised models, such as LinearRegression for linear regression, RandomForest for random forest classification and regression, and KMeans for clustering. It also provides tools for hyperparameter tuning and model selection, such as CrossValidator and TrainValidationSplit. Model deployment and evaluation: Spark ML provides a set of tools for deploying and evaluating machine learning models, such as Pipeline for building and deploying end-to-end ML pipelines, and RegressionEvaluator and MulticlassClassificationEvaluator for evaluating the performance of regression and classification models, respectively. Spark ML leverages the distributed computing capabilities of Spark to provide scalable and efficient machine learning pipelines that can process large datasets in parallel across a cluster of nodes. It also provides integration with other Spark components such as Spark SQL and Spark Streaming, enabling the seamless integration of machine learning pipelines with other data processing and analysis tasks.

1 Introduction

The rise of social media platforms has changed the way people communicate, share information, and engage with each other. While social media has brought several benefits, it has also created new challenges such as the proliferation of fake profiles, which are often used for malicious purposes. These fake profiles can cause a range of problems, including spreading misinformation, spamming, and cyberbullying. Therefore, there is a growing need to develop effective techniques for detecting and identifying fake profiles on social media platforms. One approach to address this challenge is to use big data analytics. With the explosive growth of social media data, big data analytics techniques can help analyze vast amounts of data generated by social media platforms, identify patterns, and make predictions. Big data analytics can leverage machine learning algorithms to identify features and patterns that are unique to fake profiles, enabling the creation of models that can detect and classify fake profiles in real-time. The use of big data analytics for fake profile recognition has significant implications for social media platforms, as it can help them protect their users from malicious activities. It can also help law enforcement agencies to track down and prosecute offenders who use fake profiles for illegal activities. In this paper, we will explore the use of big data analytics for fake profile recognition, including the challenges and opportunities of using this approach. We will also discuss the current state of the art in fake profile detection, and the potential for future research in this area.

2 Literature Survey

Online social media platforms have become a popular way of communication among individuals and organizations. The size of the audience commanded by an entity on these platforms is a critical measure of its popularity. However, the presence of fake profiles can bias the audience, affecting the entity's popularity. Therefore, there is a need to detect and remove such fake profiles. In this seminar report, we will discuss some related works for prediction and fake profile recognition, including image recognition techniques. Gurajala et al. (2015): In their work, Gurajala et al. studied and analyzed 62M available public profiles on Twitter and identified automatically generated fake profiles. They used the algorithm of pattern matching on screen names with an analysis tweet update time, which helped detect a reliable subset of fake user accounts. A ground truth data set analysis of profile creation time and URL of these fake accounts revealed distinct behavior. The combination of this scheme with established social graph analysis allowed time-efficient detection of fake profiles. However, it only identified a relatively small percentage of fake accounts. Guo and Zou (2017): Guo and Zou proposed a new method to retrieve images from databases, which extracts color information and uses twice clustering to help retrieve images faster and more efficiently. The model proposed by the authors is faster and efficient in visual angles and details as compared to traditional algorithms through big data framework Spark. Results of experiments conducted show that the new proposed model is faster and efficient in visual angles and details as compared to traditional algorithms. However, this paper isn't efficient enough in terms of retrieval efficiency and can be improved more. This paper needs more efficiency, and there are gaps in cluster performance, which can be improved more.

3 Methods and Approaches

Our proposed solution for fake profile prevention involves the use of Spark-based project, which is trained on 70% of the total added profile data, and then it attempts to predict the remaining 30% of data using the random forest model. Our project is based on six steps, which are described below: Reading data set from CSV: The first step is to read the data set from the CSV files. We have two CSV files, one containing the original user data, and the other containing the fake user data. We concatenate and randomize these files to create our training and testing data sets. Feature engineering: Feature engineering is an important step in the data pre-processing phase. It involves selecting the most relevant features for our model, such as age, gender, location, and other attributes of a Facebook profile. We also perform data cleaning and normalization to ensure that our data is consistent and accurate. Training data using random forest: After feature engineering, we train our data set using the random forest algorithm. Random forest is a powerful machine learning algorithm that can handle complex data sets with high accuracy. We use the MLlib library in Spark to implement the random forest model. Plotting learning curve: The learning curve is a plot of the training and testing error rates as a function of the number of training examples. It helps us to visualize how well our model is performing and whether it is overfitting or underfitting the data. We plot the learning curve to optimize our model and prevent overfitting. Plotting confusion matrix: The confusion matrix is a table that shows the true positive, false positive, true negative, and false negative predictions of our model. It helps us to evaluate the performance of our model and identify any issues with false positives or false negatives. Plotting ROC curve: The receiver operating characteristic (ROC) curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds. It helps us to evaluate the performance of our model and identify the optimal threshold for classification.

4 Data set Details

The first step in any machine learning project is to obtain the necessary data for training and testing the model. In this project, the team has gathered publicly available data of 4000+ Facebook profiles, of which less than half were identified as fake profiles. The data was obtained from various sources and was manually curated to ensure its quality.

The data was then processed and split into two separate CSV files, one containing the data for the original (real) Facebook profiles, and the other containing the data for the fake Facebook profiles. This was done to ensure that the team had separate data sets to work with for each type of profile.

To maximize the efficiency of the training and testing data set, the team concatenated and randomized the two CSV files. This was done to ensure that the data set was balanced and representative of the overall population of

Facebook profiles.

The final data set contains a mix of real and fake Facebook profiles, with each profile containing various features such as profile picture, cover photo, user name, location, education, work history, and more. The data set is well-suited for training and testing a machine learning model for detecting fake Facebook profiles.

Overall, the team has done an excellent job of gathering and processing the data for this project. The data set is diverse, representative, and well-organized, which will make it easier to build an accurate and reliable machine learning model for detecting fake Facebook profiles.

5 0 Results

Result The result of our Spark ML-based solution for fake profile detection shows a classification accuracy score of 0.94, which translates to a 94% accuracy rate. This indicates that our proposed approach is highly effective in detecting fake profiles on social media platforms. By using machine learning algorithms and Spark ML-based approach, we were able to achieve a high accuracy rate while also minimizing false positive rates to only 6%. These results are highly promising and demonstrate the potential of using advanced technologies to tackle the problem of fake profiles and online fraud. The success of our approach could potentially be applied to other areas of online security and fraud detection, opening up new avenues for research and development in this field.

6 Future Work

Explain the work to be done further.

7 Conclusion

Based on the comparison of previously used techniques and our proposed solution, we can conclude that Spark ML-based approach outperforms the other techniques. Our proposed solution has an accuracy of 94% and a false positive rate of 6%, which is better than the accuracy and false positive rates of Linear Regression, Random Forest, and Decision Tree. This indicates that our proposed solution can effectively prevent fake profiles on social media platforms. Furthermore, our proposed solution is based on a six-step process, which includes reading data set from CSV, feature engineering, training data using random forest, plotting learning curve, plotting confusion matrix, and plotting ROC curve. These steps have been designed to maximize efficiency and accuracy of the model. Overall, our proposed solution can be a valuable tool for social media companies and individuals to identify and prevent fake profiles.

8 References

AbdelAziz, A.M., Ghany, K.K.A., Soliman, T.H.A. and Sewisy, A.A.E.M. (2020) ‘A parallel multi-objective swarm intelligence framework for big data analysis’, International Journal of Computer Applications in Technology, Vol. 63, No. 3, pp.200–212. Abdullah, Yasin, A., Awan, M.J., Shehzad, M.F. and Ashraf, M. (2020) ‘Fake news classification bimodal using convolutional neural network and long short-term memory’, International Journal of Emerging Technologies in Learning, Vol. 11, No. 5, pp.209–212. Absardi, Z.N. and Javidan, R. (2017) ‘Classification of big satellite images using hadoop clusters for land cover recognition’, Proceedings of the IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), IEEE, pp.0600–0603. Aftab, M.O., Awan, M.J., Khalid, S., Javed, R. and Shabir, H. (2021) ‘Executing spark BigDL for Leukemia detection from microscopic images using transfer learning’, Proceedings of the 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), IEEE, pp.216–220. Ahmed, H.M., Awan, M.J., Khan, N.S., Yasin, A. and Shehzad, H.M.F. (2021) ‘Sentiment analysis of online food reviews using big data analytics’, Elementary Education Online, Vol. 20, No. 2, pp.827–836. Alam, T.M. and Awan, M.J (2018) ‘Domain analysis of information extraction techniques’, International Journal of Multidisciplinary Sciences and Engineering, Vol. 9, pp.1–9. Ali, Y., Farooq, A., Alam, T.M., Farooq, M.S., Awan, M.J. and Baig, T.I. (2019) ‘Detection of Schistosomiasis factors using association rule mining’, IEEE Access, Vol. 7, pp.186108–186114. Anam, M., Hussain, M., Nadeem, M.W., Javed Awan, M., Goh, H.G. and Qadeer, S. (2021) ‘Osteoporosis prediction for trabecular bone using machine learning: a review’, Computers, Materials and Continua (CMC), Vol. 67, No. 1, pp.89–10. Awan, M.J., Mohd Rahim, M.S., Salim, N., Ismail, A.W. and Shabbir, H. (2019) ‘Acceleration of knee MRI cancellous bone classification on google colaboratory using convolutional neural network’, International Journal of Advanced Trends in Computer Science and Engineering, Vol. 8, pp.83–88. Awan, M.J., Rahim, M.S.M., Nobanee, H., Munawar, A., Yasin, A. and Zain, A.M. (2021a) ‘Social media and stock market prediction: a big data approach’, Computers, Materials and Continua, Vol. 67, No. 2, pp.2569–2583. Awan, M.J., Rahim, M.S.M., Nobanee, H., Yasin, A., Khalaf, O.I. and Ishfaq, U. (2021b) ‘A big data approach to black Friday sales’, Intelligent Automation and Soft Computing, Vol. 27, No. 3, pp.785–797.