Detecting Fake Job Descriptions: An NLP Approach using PySpark MLlib

Project Proposal

Course Name:

Mining Massive Datasets

Group Members:

Abdullah Waseem	20L-1165
Fahad Waseem	20L-1134

National University Of Computer and Emerging Sciences
Department of Computer Science
Lahore, Pakistan

1. Project Overview

Our project focuses on developing an NLP algorithm using PySpark MLlib to automatically flag suspicious job posts for review, based on a dataset containing 18,000 job descriptions, including approximately 800 fake ones. By leveraging machine learning techniques, we aim to create a classification model capable of distinguishing between genuine and fraudulent job descriptions, thus aiding in the detection of potentially deceptive postings.

2. Mining of Massive Dataset Techniques

In this project, we aim to tackle the challenge of detecting fraudulent job postings by employing a comprehensive approach that combines advanced natural language processing (NLP) techniques with state-of-the-art machine learning algorithms. Leveraging PySpark's MLlib, we will harness the power of distributed computing to efficiently process the massive dataset comprising approximately 18,000 job descriptions, with around 800 flagged as fraudulent.

To transform the raw textual data into numerical feature vectors, we will utilize both Count Vectors and TF-IDF (Term Frequency-Inverse Document Frequency) representations. These techniques enable us to capture the frequency of words in each job description, effectively encoding the textual information in a format suitable for machine learning algorithms. Additionally, we will explore the use of Word2Vec, a powerful embedding technique that learns distributed representations of words in a continuous vector space. By training Word2Vec models on our text data, we can generate dense vector embeddings for words, capturing semantic relationships based on contextual usage within the text.

With our feature vectors in hand, we will proceed to implement various text classification algorithms, including but not limited to Logistic Regression, Naive Bayes, Random Forest, Gradient Boosted Trees, Linear Support Vector Machines (SVM), Decision Trees, and Multilayer Perceptron Classifier. These algorithms will learn patterns from the vectorized textual features and classify job descriptions as either genuine or suspicious/fraudulent.

However, building effective classification models is only part of the process. We must also rigorously evaluate and select the best-performing models. To achieve this, we will employ a range of evaluation metrics such as accuracy, precision, recall, and F1-score. Moreover, we will utilize techniques like cross-validation and hyperparameter tuning to optimize model performance and prevent overfitting.

By leveraging PySpark's MLlib for distributed NLP and machine learning, we aim to develop robust and scalable models capable of accurately detecting fraudulent job postings.

3. Related Work

The referenced research paper proposes using machine learning classifiers to identify fraudulent job postings online. By analyzing historical data, the system trains models to differentiate between genuine and deceptive job listings. Our approach differs as we plan to implement MLlib and NLP techniques using Apache Spark. Leveraging Spark's scalability, we aim to efficiently process large datasets, while NLP enhances our model's ability to detect subtle fraudulent patterns. This integration will contribute to the reliability of online job markets by effectively flagging suspicious postings.[1]

[1] Khandagale, P., Utekar, A., Dhonde, A., & Karve, S. S. (1822). Fake Job Detection using Machine Learning. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 10 (5), 1826.

4. Dataset Details

The dataset comprises approximately 18,000 job descriptions, with approximately 800 labeled as fraudulent. Each entry contains textual information describing the job roles, responsibilities, and requirements, along with meta-information such as job title, location, department, salary range, company profile, employment type, required experience, required education, industry, function, and whether the job is fraudulent or not.

The textual descriptions provide insights into the roles and responsibilities associated with each job, while the meta-information offers contextual details such as the location of the job, the industry it belongs to, and the level of experience required. This diverse dataset is suitable for building classification models aimed at identifying fraudulent job postings based on the provided textual and meta-information features.

The dataset encompasses various industries and job functions, offering a wide range of scenarios for training and evaluating classification algorithms. Researchers and practitioners can utilize this dataset to develop and assess machine learning models that can effectively detect fraudulent job postings, thereby aiding in the prevention of fraudulent activities in online recruitment platforms.

The Dataset is avaliable on Kaggle at
https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction