Data Governance Review:

Fake Job Posting Detection

Date of Review: March 13, 2024

Overview:

The data governance review for the Fake Job Posting Detection project aims to assess the quality, security, compliance, documentation, and lifecycle management of the data used in the project. The review will ensure that the data meets predefined standards and requirements, aligning with the project's objectives.

1. Data Quality: Evaluate the quality of the data used in the project, ensuring it meets predefined standards and requirements.

1.1. Data Accuracy:

1.1.1. Assessment : The data accuracy appears to be high, with minimal observed errors or inconsistencies. No obvious outliers for numeric data and inconsistencies found for textual data found. The dataset has also been trusted by many kaggle users.

1.1.2. Action Items : No immediate action required.

1.2. Completeness:

1.2.1. Assessment : The assessment of data completeness based on missing values in the entire DataFrame reveals varying levels of completeness across different features. The following features exhibit missing values:

- location: 346 missing values
- department: 11547 missing values
- salary_range: 15012 missing values
- company_profile: 3308 missing values
- description: 1 missing value
- requirements: 2695 missing values
- benefits: 7210 missing values
- employment_type: 3471 missing values
- required_experience: 7050 missing values
- required_education: 8105 missing values
- industry: 4903 missing values
- function: 6455 missing values

1.2.2. Action Items : Develop procedures to handle missing data, including imputation techniques or considering alternative data sources to fill in missing information where feasible. Additionally,

prioritize data collection efforts for critical features with significant missing values to ensure comprehensive dataset coverage.

1.3. Consistency:

1.3.1. Assessment : The presence of strange characters like "Œ†ŒõŒóŒ°ŒüŒ¶ŒüŒ°ŒôŒöŒó" in some cells indicates potential data consistency issues, possibly due to encoding or formatting errors.

1.3.2. Action Items : Conduct a thorough review of the data to identify and rectify any encoding or formatting inconsistencies. This may involve standardizing encoding formats or utilizing appropriate data cleaning techniques to ensure consistency and readability throughout the dataset.

1.4. Timeliness:

1.4.1. Assessment : Timeliness is not applicable in the context of this project as the dataset does not inherently provide temporal information or timestamps related to data collection or updates. However, it's important to note that the absence of timeliness data does not affect the potential of the project, as the focus is primarily on the content and quality of the job postings rather than the timing of their creation or updates.

1.4.2. Action Items : Since timeliness concerns are not applicable to the dataset, no specific action items are required in this regard.

2. Data Security: Assess the security measures in place to protect sensitive and confidential data.

2.1. Access Controls:

2.1.1. Assessment : Access controls are not applicable in this scenario as the dataset used for the project is publicly available and does not require restricted access.

2.1.2. Action Items : No specific action items are required for access controls since the dataset is publicly accessible.

2.2. Encryption:

2.2.1. Assessment : Encryption is not applicable as the dataset is publicly available and does not contain sensitive or confidential information that requires encryption.

2.2.2. Action Items : No action items are necessary for encryption as the dataset does not contain sensitive data.

2.3. Data Masking/Anonymization:

2.3.1. Assessment : Data masking/anonymization is not applicable since the dataset used for the project is publicly available and does not contain personally identifiable information or sensitive data that requires anonymization.

2.3.2. Action Items : No specific action items are necessary for data masking/anonymization as the dataset does not contain sensitive information.

3. Data Compliance: Review compliance with relevant data protection regulations and internal policies.

3.1. Regulatory Compliance (e.g., GDPR, HIPAA):

3.1.1. Assessment : The dataset used in the project does not contain personally identifiable information (PII) or sensitive data regulated by GDPR or HIPAA.

3.1.2. Action Items : Since the dataset does not contain data subject to GDPR or HIPAA regulations, no specific action items are required for compliance.

3.2. Internal Policies:

3.2.1. Assessment : Internal policies regarding data handling and privacy protection are not directly applicable to this project as it utilizes a publicly available dataset.

3.2.2. Action Items : Ensure that any future datasets or data sources used in the project adhere to internal policies regarding data privacy and security, if applicable.

4. Data Documentation: Evaluate the completeness and accuracy of data documentation.

4.1. Metadata:

4.1.1.1. Assessment : The dataset includes metadata such as column names and data types, providing basic documentation.

4.1.1.2. Action Items : Potentially enhance metadata documentation by providing additional details such as the source of the data, date of last update, and any transformations applied.

4.1.2. Data Catalog:

4.1.2.1. Assessment : A comprehensive data catalog is not applicable since no additional datasets are being used or collected.

4.1.2.2. Action Items : No immediate actions needed.

4.1.3. Data Lineage:

4.1.3.1. Assessment : Data lineage information, tracing the origin and movement of data, is not provided for the dataset.

4.1.3.2. Action Items : Establish a data lineage process to track the flow of data from its sources through various transformations to its final use in the project, ensuring transparency and reproducibility.

5. Data Lifecycle Management: Assess how data is handled throughout its lifecycle, from creation to archival.

5.1.1. Data Creation and Collection:

5.1.1.1. Assessment : The data was obtained from a public dataset on kaggle, indicating that it was created and collected by a third party.

5.1.1.2. Action Items :

- Ensure proper citation to the original source of the dataset.
- Document the date of dataset retrieval and any relevant information regarding its creation.

5.1.2. Data Storage and Retention:

5.1.2.1. Assessment :

- The dataset is currently stored in a local system.
- There is no specific retention policy mentioned as the dataset is static and does not require frequent updates.

5.1.2.2. Action Items : No immediate actions needed.

5.1.3. Data Archiving and Deletion:

5.1.3.1. Assessment : There is no current plan for data archiving or deletion due to the public nature of the dataset.

5.1.3.2. Action Items : No immediate actions needed.

6. Recommendations and Action Plan: Summarize key findings and provide recommendations for improvements in data governance.

6.1.1. Immediate Actions:

6.1.1.1.1. Action 1 : Implement data validation checks to ensure the accuracy and completeness of critical fields, such as job title, location, and company information.

6.1.1.1.2. Action 2 : Develop a standardized process for handling missing data and removing duplicates, including imputation techniques where applicable, to improve data completeness.

6.1.2. Long-term Improvements:

6.1.2.1.1. Recommendation 1 : Establish a comprehensive data governance framework outlining roles, responsibilities, and procedures for data management and quality assurance.

6.1.2.1.2. Recommendation 2 : Implement regular audits and reviews of the dataset to identify and address any emerging data quality issues, ensuring ongoing data integrity and reliability.

7. Follow-up Plan: Outline the plan for implementing the recommended actions and schedule follow-up reviews.

7.1. Timeline:

7.1.1. Timeline for Immediate Actions

- Implement data validation checks: Within the next two weeks.

- Develop a standardized process for handling missing data: Within the next month.

7.1.2. Timeline for Long-term Improvements: By the end of the project