

Homework

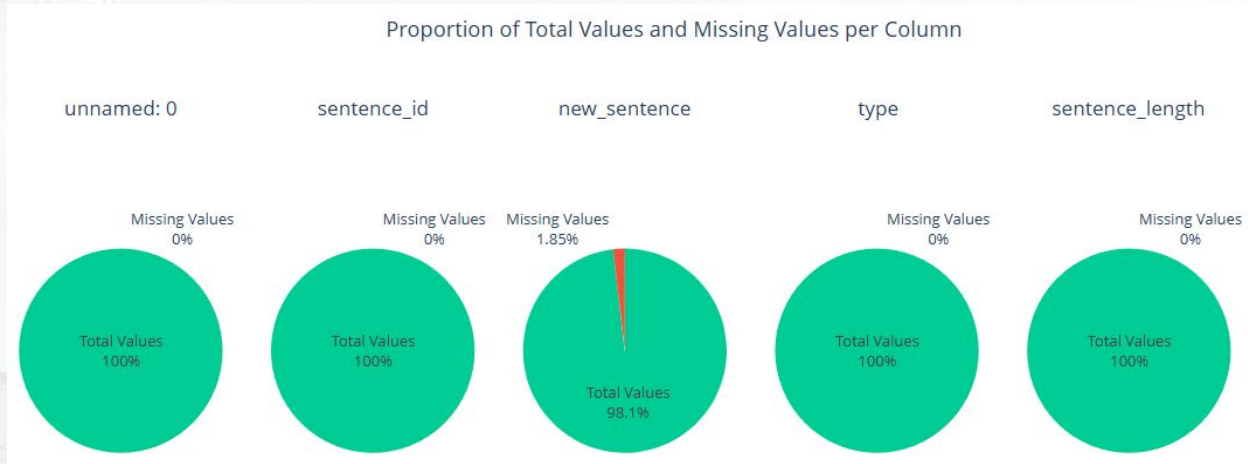
Stage-1

Presented by NautixTech

Descriptive Statistic

	unnamed: 0	sentence_id	new_sentence	type	sentence_length
count	60115.000000	60115	59002	60115	60115.000000
unique	NaN	59704	57991	6	NaN
top	NaN	HONREQ4583	his is a dummy block of text And this is repre...	Responsibility	NaN
freq	NaN	2	852	15561	NaN
mean	30057.000000	NaN	NaN	NaN	88.093454
std	17353.850053	NaN	NaN	NaN	67.456955
min	0.000000	NaN	NaN	NaN	0.000000
25%	15028.500000	NaN	NaN	NaN	50.000000
50%	30057.000000	NaN	NaN	NaN	74.000000
75%	45085.500000	NaN	NaN	NaN	112.000000
max	60114.000000	NaN	NaN	NaN	5419.000000

1. In the **sentence_id** column, the most frequently occurring value is HONREQ4583, appearing 2 times.
2. In the **new_sentence** column, the most frequently occurring sentence is "this is a dummy block of text And this is representative of example sentences," appearing 852 times, and this column has 1153 missing values.
3. In the **type** column, there are 6 categories, with the most frequently occurring category being Responsibility, appearing 15,561 times.

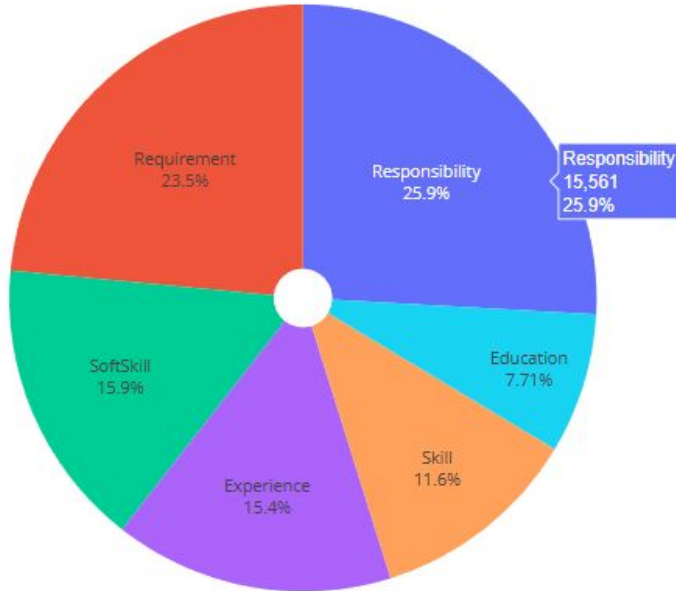


Based on the visualization above, the column with missing values is **new_sentence**, with 1.85% missing. Since the amount of missing values is less than 20%, they will be removed.

Univariate Analysis

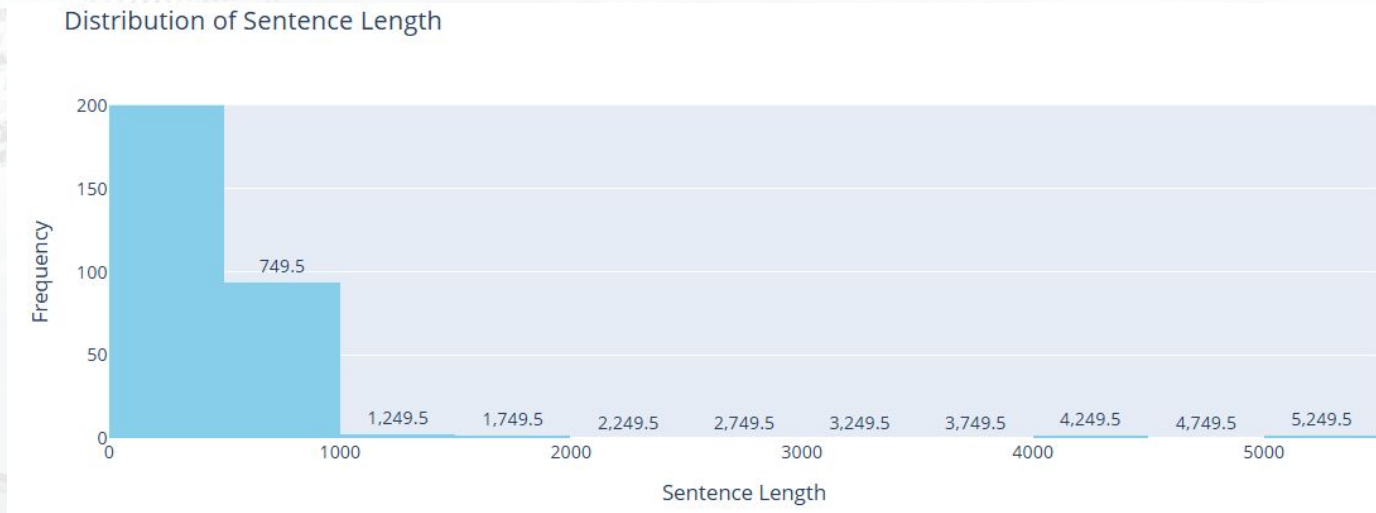
Proportion of Each Category

Proportion of Each Type



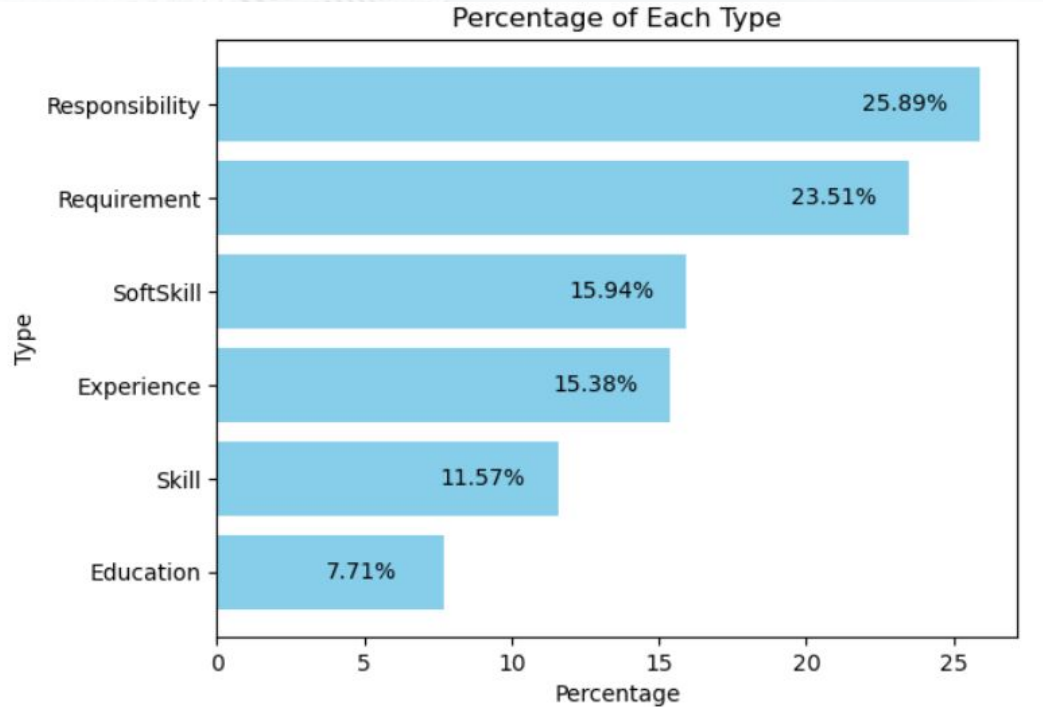
Based on the results, the most frequently occurring category is Responsibility, accounting for 25.9%. The categories "Skill," "Education," and "Experience" have smaller proportions compared to the top three categories. This may indicate that these aspects are considered supplementary or not heavily emphasized in the dataset.

Distribution of Sentence Length



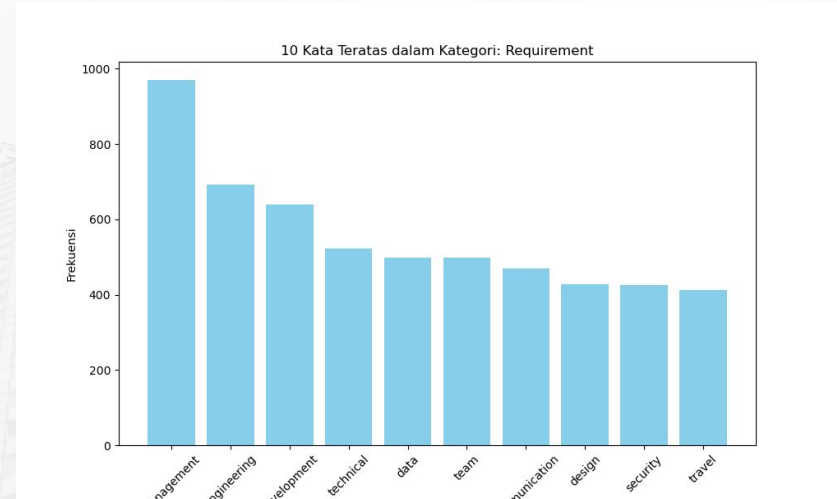
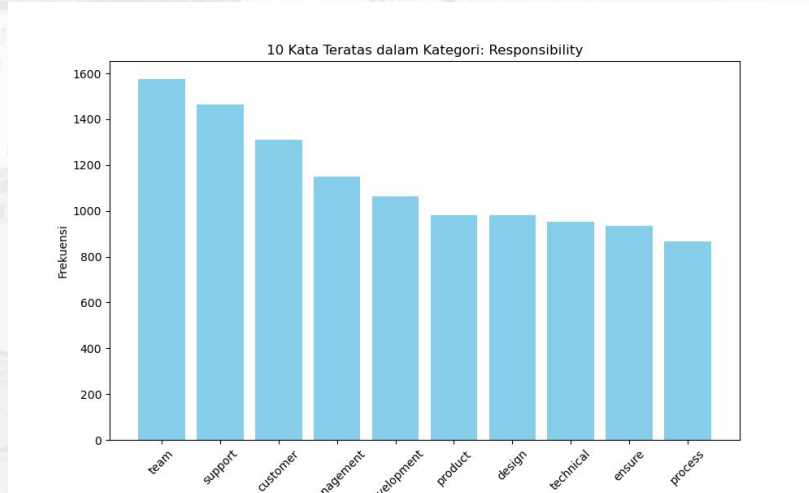
Most sentences have a relatively short length, falling within the lowest sentence length range, which is below 1000 characters. This indicates that the text in the dataset tends to consist of short and concise sentences.

Imbalance Class



Based on the visualization results, there is an imbalance in the classes, although it is not extreme. The class distribution in the target variable shows a slight imbalance, with the majority class accounting for approximately 25.89% of the total data, while the minority class accounts for 7.71%.

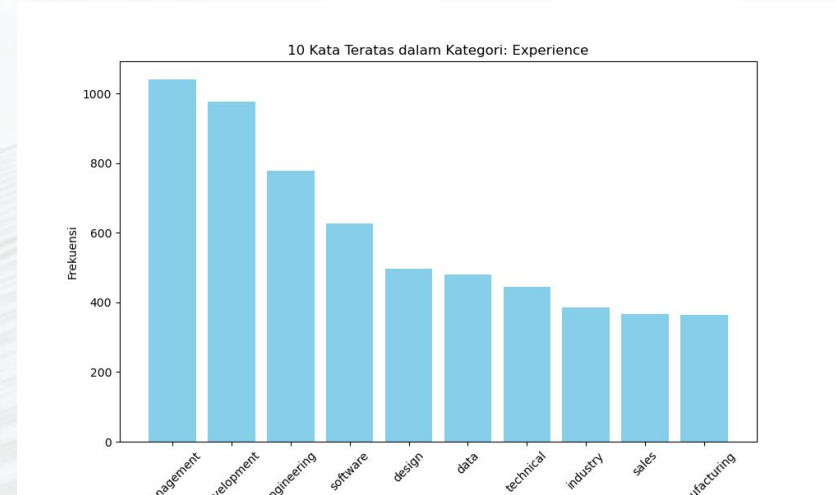
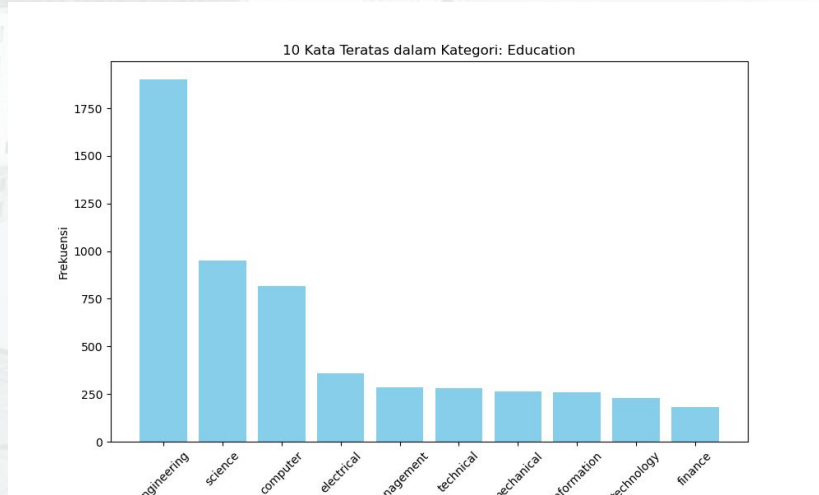
Keyword Frequency



In the **Responsibility** category, the frequently occurring words are team, support, customer, management, development, product, design, technical, ensure, and process. From these keywords, it can be concluded that this category highly emphasizes interaction patterns, result orientation, and management aspects.

Meanwhile, in the **Requirement** category, the frequently occurring words are management, engineering, development, technical, data, team, communication, design, security, and travel. From these keywords, it can be concluded that this category highly emphasizes technical qualifications and specific skills.

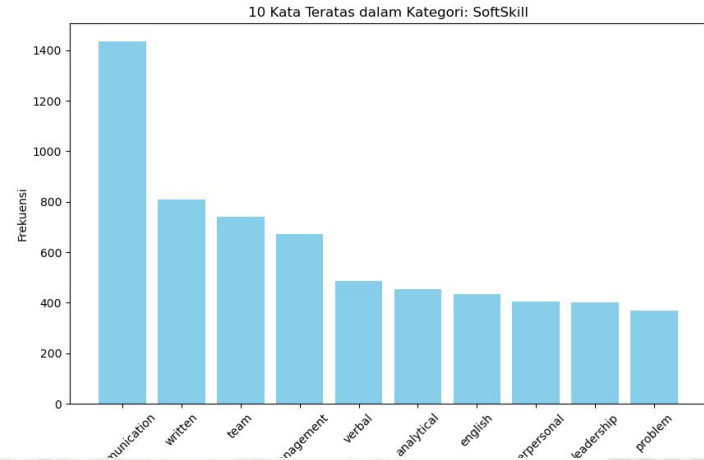
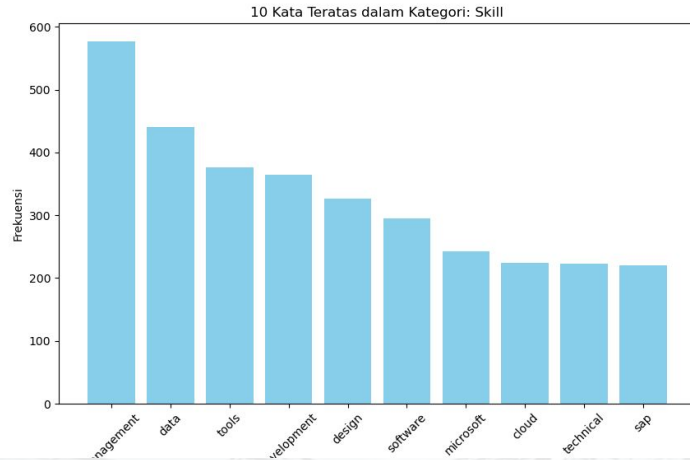
Keyword Frequency



In the **Education** category, the most frequently sought educational backgrounds are engineering, science, and computer.

Meanwhile, in the **Experience** category, the frequently occurring words are management, engineering, development, software, design, data, technical, industry, sales, and manufacturing. From these keywords, it can be concluded that this category highly emphasizes experience in the STEM fields.

Keyword Frequency

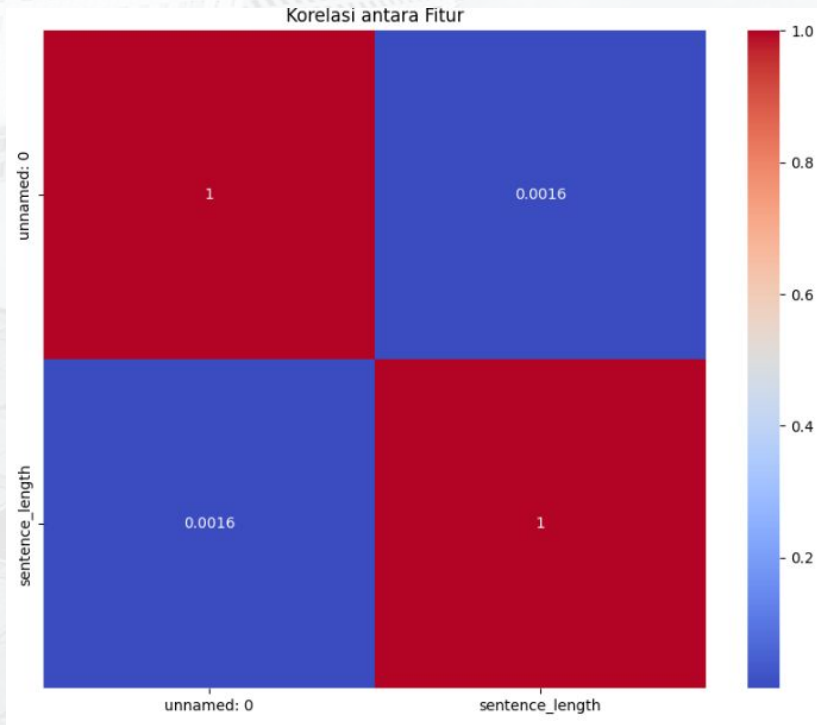


In the **Skill** category, the frequently occurring words are management, data, tools, development, design, software, microsoft, cloud, technical, and sap. From these keywords, it can be concluded that this category highly emphasizes technical and analytical skills.

Meanwhile, in the **Softskill** category, the frequently occurring words are communication, written, team, management, verbal, analytical, english, interpersonal, leadership, and problem. From these keywords, it can be concluded that this category highly emphasizes communication, leadership, and interpersonal skills.

Multivariate Analysis

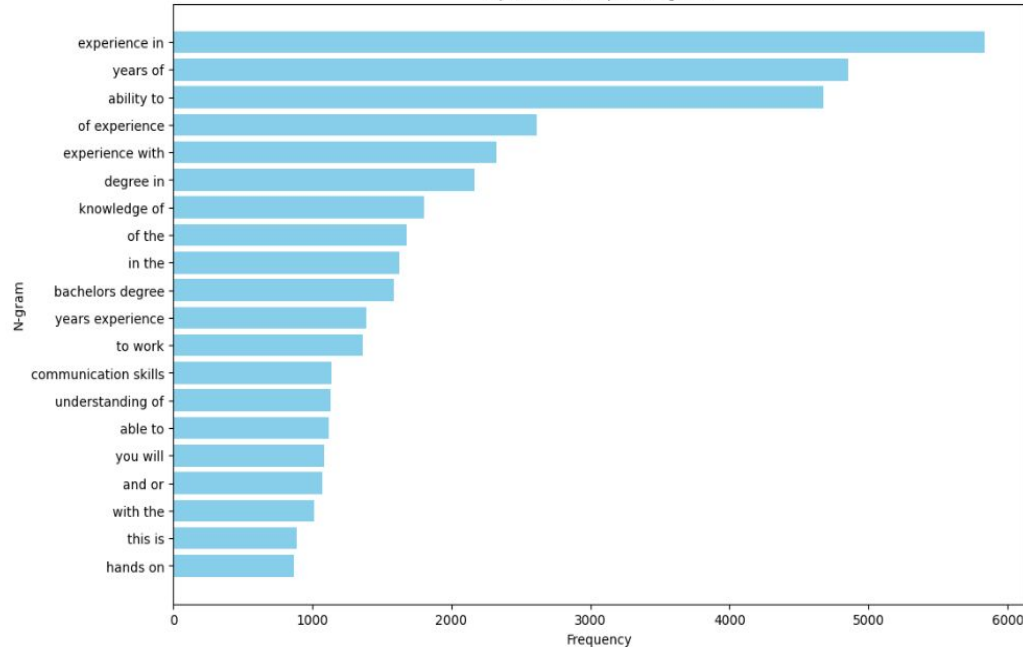
Correlation Between Features



The correlation results above show a value of 0.0016 for the correlation between sentence length and index. It can be concluded that the absence of a very strong correlation indicates that the two features are closely related and may influence each other.

Ngram Analytics

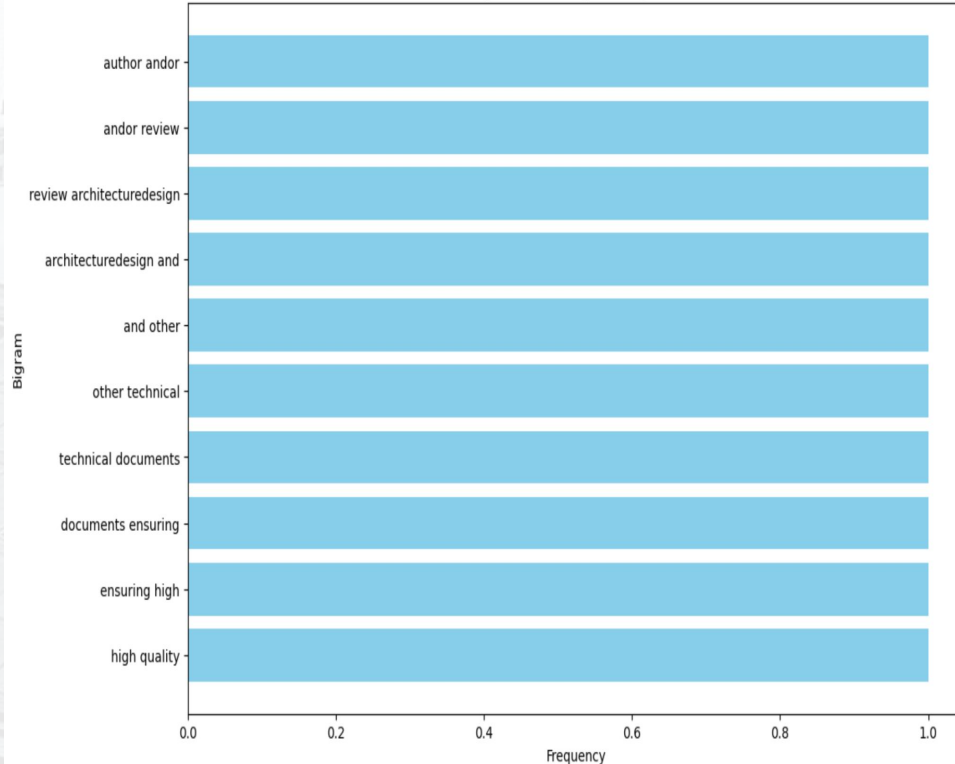
Top 20 Most Frequent N-grams



The graph shows the 20 most frequently occurring word combinations (n-grams) in a text corpus. The most frequently occurring phrases are "experience in," "years of," and "ability to," which focus on a person's experience, skills, and knowledge.

Ngram Analytics

Top 10 Most Frequent Bigrams



The chart shows the **top 10 most frequent bigrams (pairs of consecutive words)** in the dataset, such as "author andor," "andor review," and "technical documents." The frequencies are nearly identical, indicating consistent appearances across the text. These bigrams highlight a focus on technical documents, high-quality standards, and review processes, which are relevant in professional or technical contexts. This information can help identify the main themes of the text or serve as features for further analysis, such as classification or machine learning tasks.

Business Insight

1. **Responsibility Dominance:** The Responsibility category is the most frequently occurring, accounting for 25.9% of the data. This indicates a strong emphasis on roles and tasks related to team interaction, customer support, and management.
2. **Technical Qualifications:** The Requirement category highlights the importance of technical qualifications and specific skills, with frequent keywords like management, engineering, development, and technical.
3. **STEM Focus:** The Experience category shows a significant focus on experience in STEM fields, with common words including management, engineering, development, software, and design.
4. **Educational Background:** In the Education category, the most sought-after educational backgrounds are in engineering, science, and computer fields.
5. **Communication and Interpersonal Skills:** The Softskill category highlights the importance of communication, leadership, and interpersonal skills, with common words like communication, written, team, management, and verbal.

Business Recommendation

1. **Enhance Training Programs:** Develop training programs focused on improving technical qualifications and specific skills, especially in the areas highlighted in the Requirement and Skill categories.
2. **Promote STEM Education:** Encourage and support educational programs in STEM fields to align with the demand for engineering, science, and computer backgrounds.
3. **Focus on Soft Skills Development:** Implement initiatives to enhance communication, leadership, and interpersonal skills, as these are crucial in the Softskill category.
4. **Address Class Imbalance:** Consider techniques to address the slight class imbalance, such as data augmentation or resampling methods, to ensure a more balanced dataset.
5. **Tailor Recruitment Strategies:** Adjust recruitment strategies to prioritize candidates with experience in STEM fields and strong technical and analytical skills.

Things to-do

From the EDA results, data pre-processing will be continued. The possible stages are:

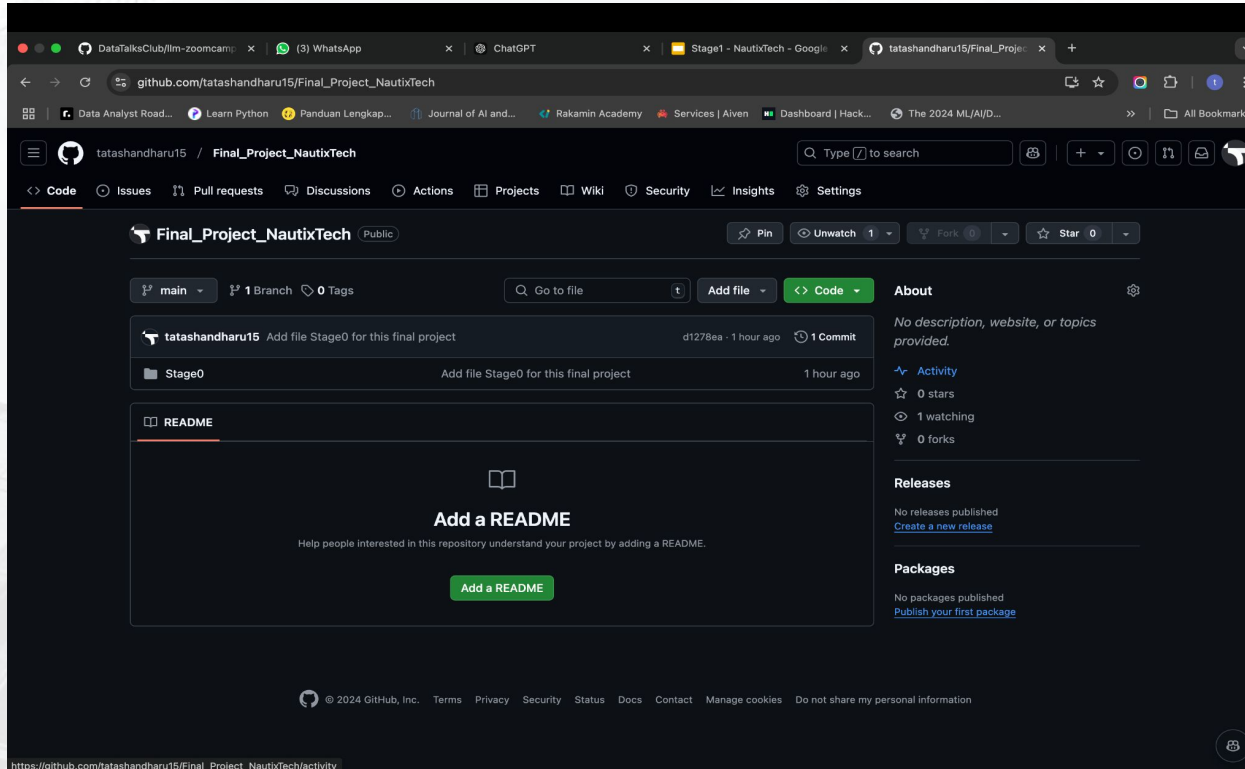
- Removing missing values because the percentage is less than 20%
- Check the imbalance class and review what method to overcome it, whether just ignoring it or using oversampling/undersampling techniques
- Cleaning text by checking for typos, or analyzing the words
- Tokenize words which are very important for NLP
- From the results of the multivariate analysis, all features will be used and one hot encoding will be carried out for the feature type



Git

Link Repository

[tatashandharu15/Final_Project_NautixTech](https://github.com/tatashandharu15/Final_Project_NautixTech)



The screenshot shows a web browser window displaying a GitHub repository page. The browser's address bar shows the URL `github.com/tatashandharu15/Final_Project_NautixTech`. The repository page has a dark theme. At the top, the repository name `Final_Project_NautixTech` is shown as public, with options to pin, unwatch (1), fork, and star (0). Below this, the main branch is `main` with 1 branch and 0 tags. A commit by `tatashandharu15` is listed, titled "Add file Stage0 for this final project", dated "d1278ea · 1 hour ago", with 1 commit. A file named `Stage0` is shown, also titled "Add file Stage0 for this final project" and dated "1 hour ago". A `README` section is present, prompting the user to "Add a README" with the text "Help people interested in this repository understand your project by adding a README." and a green "Add a README" button. On the right side, the "About" section states "No description, website, or topics provided." Below it, the "Activity" section shows 0 stars, 1 watching, and 0 forks. The "Releases" section states "No releases published" with a link to "Create a new release". The "Packages" section states "No packages published" with a link to "Publish your first package". The footer of the page includes the GitHub logo, copyright information for 2024 GitHub, Inc., and links for Terms, Privacy, Security, Status, Docs, Contact, Manage cookies, and Do not share my personal information. The URL `https://github.com/tatashandharu15/Final_Project_NautixTech/activity` is visible at the bottom left.

Thank You