Rakamin Academy

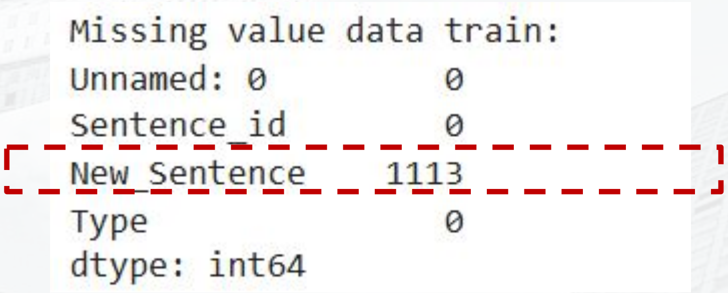
Homework

Stage-2

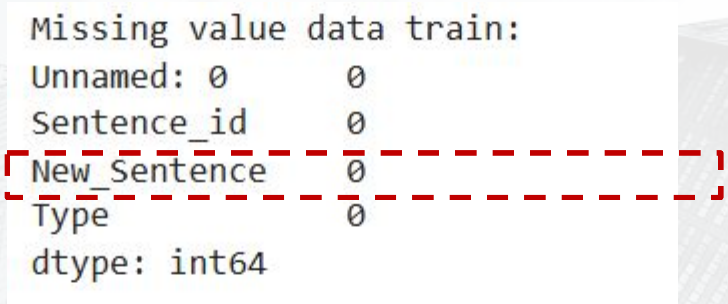
Presented by NautixTech

Data Cleansing

Missing Value

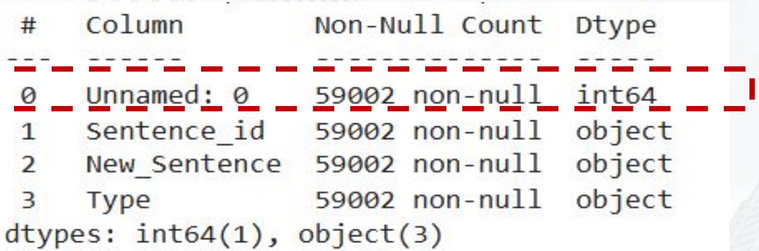


```
Missing value data train:
Unnamed: 0         0
Sentence_id        0
New_Sentence    1113
Type               0
dtype: int64
```



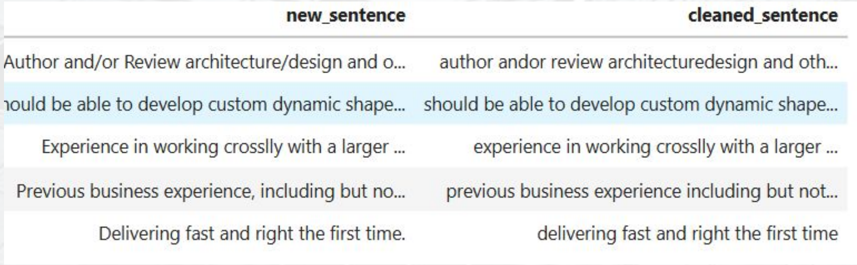
```
Missing value data train:
Unnamed: 0      0
Sentence_id     0
New_Sentence    0
Type            0
dtype: int64
```

Delete Irrelevant Columns



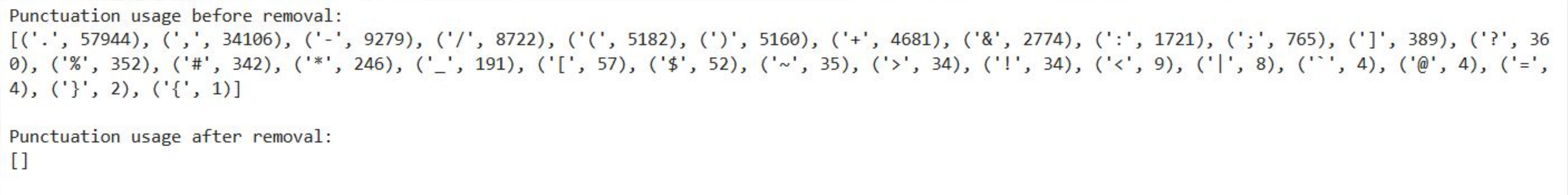
```
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Unnamed: 0     59002 non-null  int64
 1   Sentence_id    59002 non-null  object
 2   New_Sentence   59002 non-null  object
 3   Type           59002 non-null  object
dtypes: int64(1), object(3)
```

Lowercase



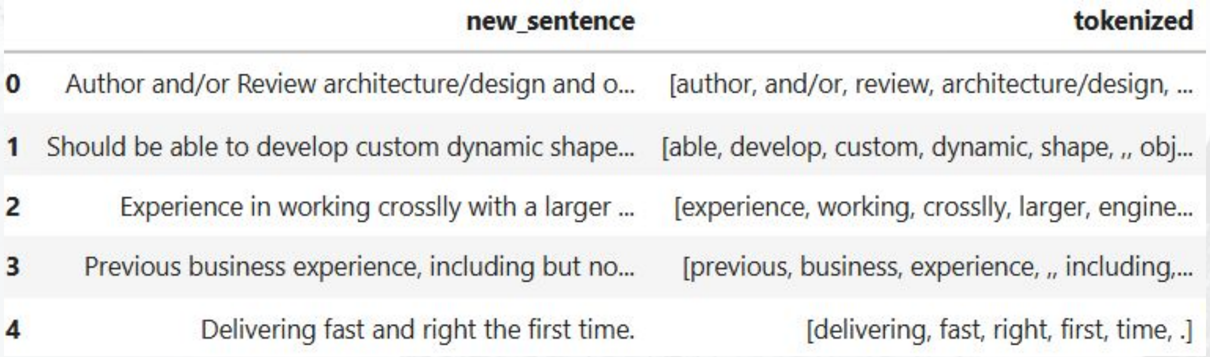
new_sentence	cleaned_sentence
Author and/or Review architecture/design and o...	author andor review architecturedesign and oth...
hould be able to develop custom dynamic shape...	should be able to develop custom dynamic shape...
Experience in working crosslly with a larger ...	experience in working crosslly with a larger ...
Previous business experience, including but no...	previous business experience including but not...
Delivering fast and right the first time.	delivering fast and right the first time

Punctuations Removal



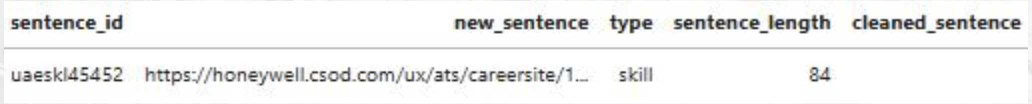
```
Punctuation usage before removal:
[('.', 57944), (',', 34106), ('-', 9279), ('/', 8722), ('(', 5182), (')', 5160), ('+', 4681), ('&', 2774), (':', 1721), (';', 765), (']', 389), ('?', 36
0), ('%', 352), ('#', 342), ('*', 246), ('_', 191), ('[', 57), ('$', 52), ('~', 35), ('>', 34), ('!', 34), ('<', 9), ('|', 8), ('`', 4), ('@', 4), ('=',
4), ('}', 2), ('{', 1)]

Punctuation usage after removal:
[]
```

Remove StepWords & Tokenization



	new_sentence	tokenized
0	Author and/or Review architecture/design and o...	[author, and/or, review, architecture/design, ...
1	Should be able to develop custom dynamic shape...	[able, develop, custom, dynamic, shape, ,, obj...
2	Experience in working crosslly with a larger ...	[experience, working, crosslly, larger, engine...
3	Previous business experience, including but no...	[previous, business, experience, ,, including,...
4	Delivering fast and right the first time.	[delivering, fast, right, first, time, .]

Remove URL Links



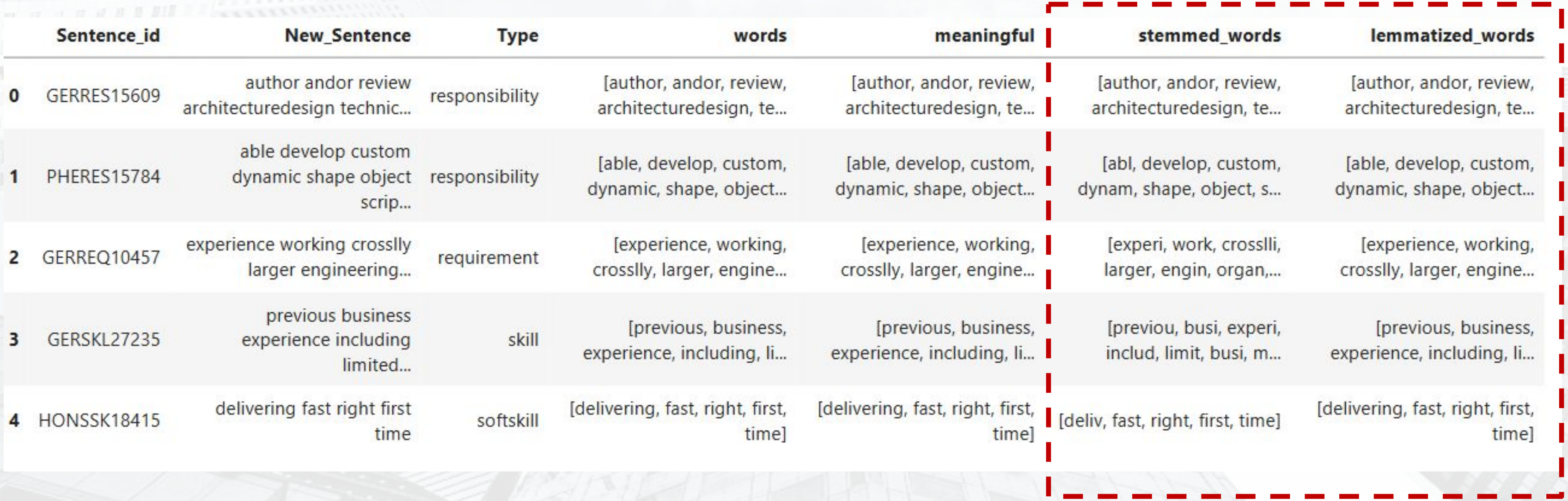
sentence_id	new_sentence	type	sentence_length	cleaned_sentence
uaeskl45452	https://honeywell.csod.com/ux/ats/careersite/1...	skill	84	

In this project, the data cleaning stages carried out include:

1. Handle missing value
   There are 1,113 empty rows in the new_statement column. In this case, the empty rows will be deleted.
2. Delete irrelevant columns
   There is an irrelevant column named Unnamed: 0. Therefore, this column will be deleted.
3. Changing Text Format in the new_statement Column
   The sentences in the new_statement column will be converted to lowercase.
4. Remove Irrelevant Punctuation
   There are some punctuation marks that will affect the modeling results. These punctuation marks will be removed.
5. Remove StepWords dan Tekonization
   The sentences in the new_statement column will be tokenized so that they are separated into individual words. Then, stopword removal will be performed by extracting important words from the tokenized results.
6. Remove URL Links
   There is one row that contains a URL. This URL will be removed, making the row empty. The empty row will be deleted as it becomes a row with no value.

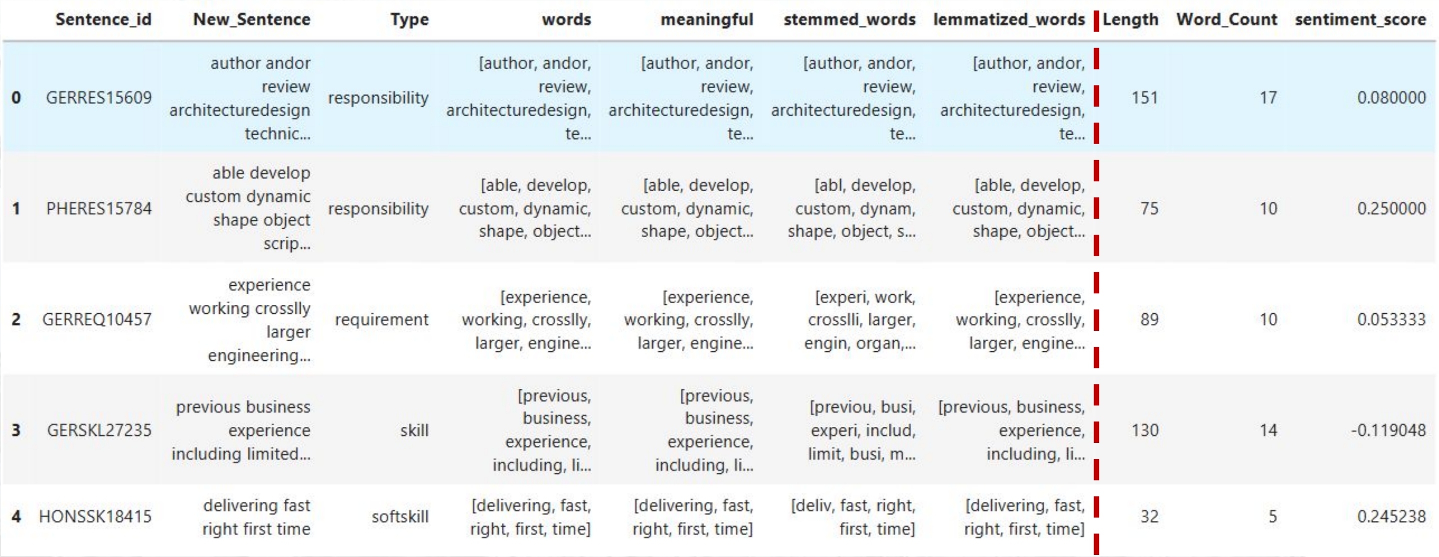
Feature Engineering

Stemming & Lemmatization



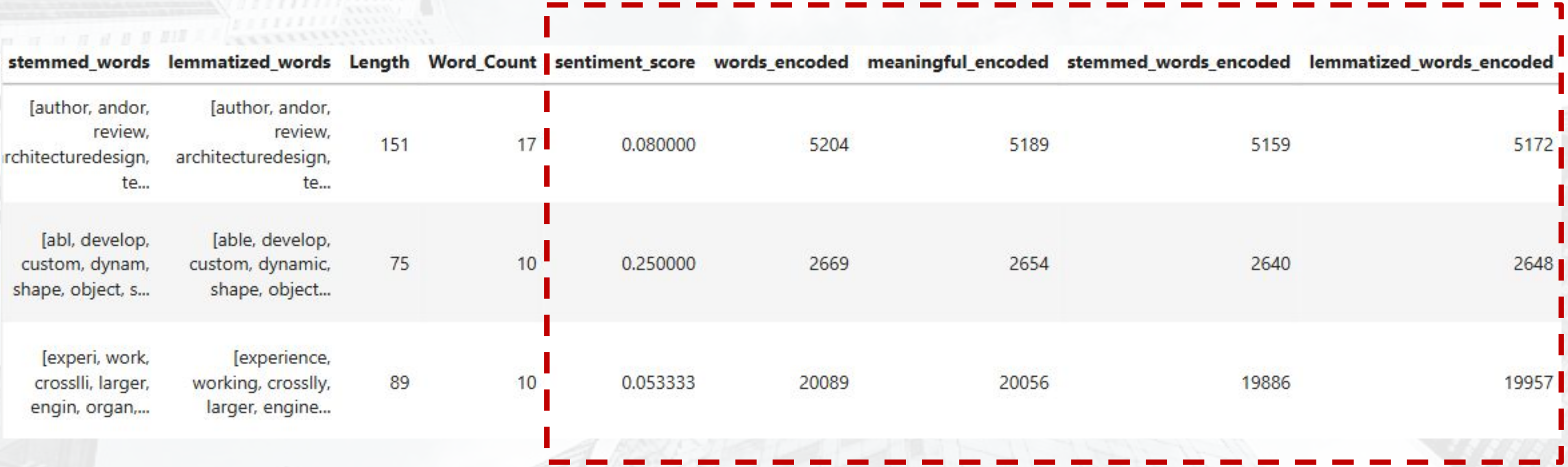
	Sentence_id	New_Sentence	Type	words	meaningful	stemmed_words	lemmatized_words
0	GERRES15609	author andor review architecturedesign technic...	responsibility	[author, andor, review, architecturedesign, te...	[author, andor, review, architecturedesign, te...	[author, andor, review, architecturedesign, te...	[author, andor, review, architecturedesign, te...
1	PHERES15784	able develop custom dynamic shape object scrip...	responsibility	[able, develop, custom, dynamic, shape, object...	[able, develop, custom, dynamic, shape, object...	[abl, develop, custom, dynam, shape, object, s...	[able, develop, custom, dynamic, shape, object...
2	GERREQ10457	experience working crosslly larger engineering...	requirement	[experience, working, crosslly, larger, engine...	[experience, working, crosslly, larger, engine...	[experi, work, crosslli, larger, engin, organ,...	[experience, working, crosslly, larger, engine...
3	GERSKL27235	previous business experience including limited...	skill	[previous, business, experience, including, li...	[previous, business, experience, including, li...	[previou, busi, experi, includ, limit, busi, m...	[previous, business, experience, including, li...
4	HONSSK18415	delivering fast right first time	softskill	[delivering, fast, right, first, time]	[delivering, fast, right, first, time]	[deliv, fast, right, first, time]	[delivering, fast, right, first, time]

Feature Engineering



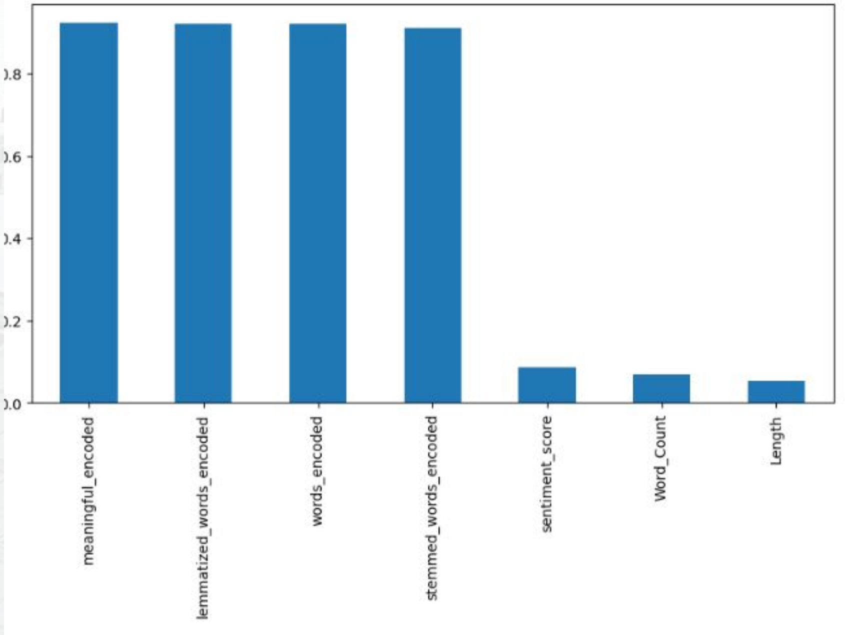
	Sentence_id	New_Sentence	Type	words	meaningful	stemmed_words	lemmatized_words	Length	Word_Count	sentiment_score
0	GERRES15609	author andor review architecturedesign technic...	responsibility	[author, andor, review, architecturedesign, te...	[author, andor, review, architecturedesign, te...	[author, andor, review, architecturedesign, te...	[author, andor, review, architecturedesign, te...	151	17	0.080000
1	PHERES15784	able develop custom dynamic shape object scrip...	responsibility	[able, develop, custom, dynamic, shape, object...	[able, develop, custom, dynamic, shape, object...	[abl, develop, custom, dynam, shape, object, s...	[able, develop, custom, dynamic, shape, object...	75	10	0.250000
2	GERREQ10457	experience working crosslly larger engineering...	requirement	[experience, working, crosslly, larger, engine...	[experience, working, crosslly, larger, engine...	[experi, work, crosslli, larger, engin, organ,...	[experience, working, crosslly, larger, engine...	89	10	0.053333
3	GERSKL27235	previous business experience including limited...	skill	[previous, business, experience, including, li...	[previous, business, experience, including, li...	[previou, busi, experi, includ, limit, busi, m...	[previous, business, experience, including, li...	130	14	-0.119048
4	HONSSK18415	delivering fast right first time	softskill	[delivering, fast, right, first, time]	[delivering, fast, right, first, time]	[deliv, fast, right, first, time]	[delivering, fast, right, first, time]	32	5	0.245238

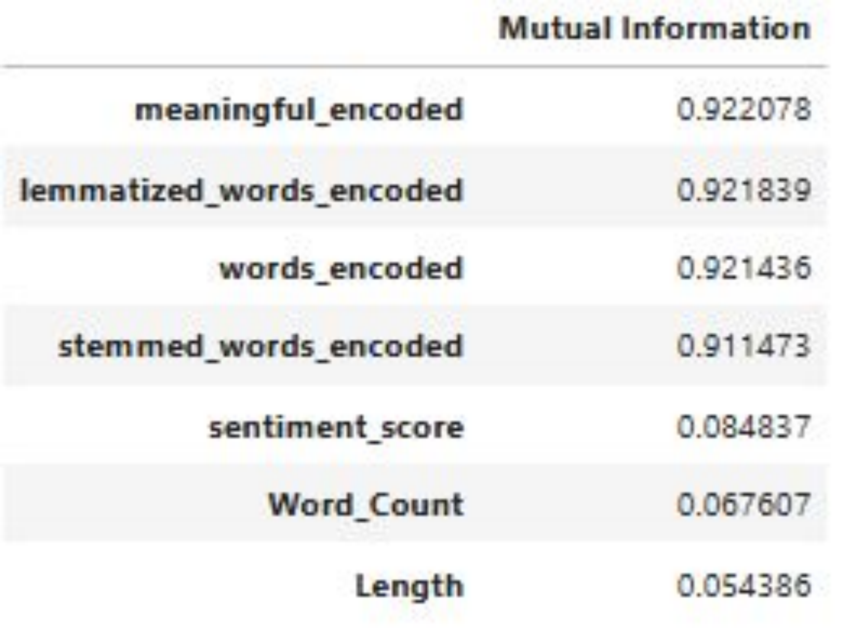
Evaluation Feature



stemmed_words	lemmatized_words	Length	Word_Count	sentiment_score	words_encoded	meaningful_encoded	stemmed_words_encoded	lemmatized_words_encoded
[author, andor, review, rchitecturedesign, te...	[author, andor, review, architecturedesign, te...	151	17	0.080000	5204	5189	5159	5172
[abl, develop, custom, dynam, shape, object, s...	[able, develop, custom, dynamic, shape, object...	75	10	0.250000	2669	2654	2640	2648
[experi, work, crosslli, larger, engin, organ,...	[experience, working, crosslly, larger, engine...	89	10	0.053333	20089	20056	19886	19957

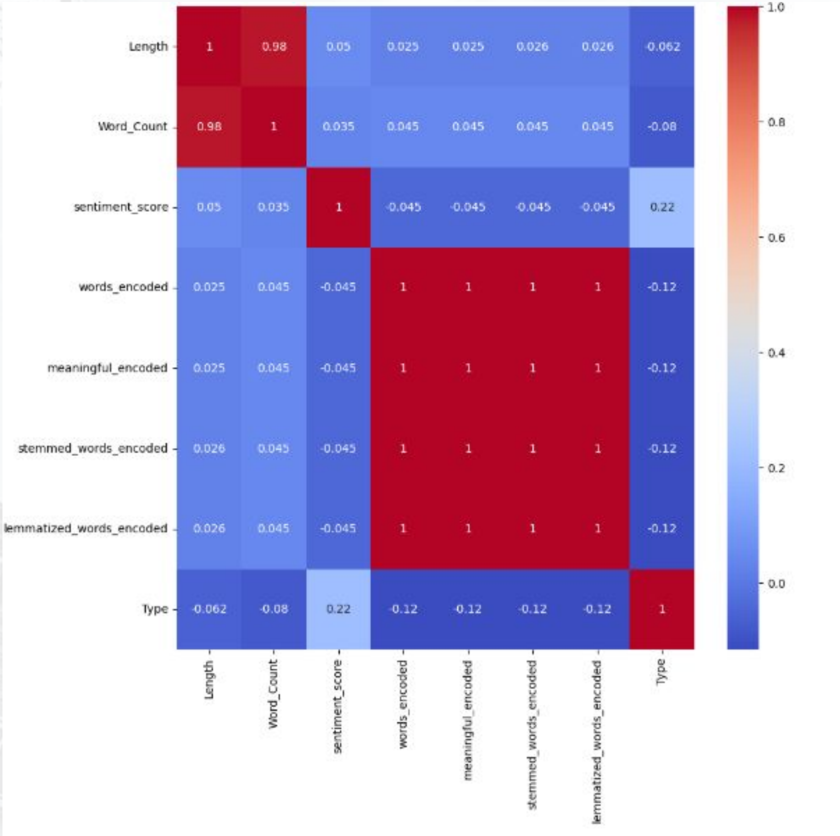
Evaluation Feature





	Mutual Information
meaningful_encoded	0.922078
lemmatized_words_encoded	0.921839
words_encoded	0.921436
stemmed_words_encoded	0.911473
sentiment_score	0.084837
Word_Count	0.067607
Length	0.054386

Evaluation Feature



A. Stemming & Lemmatization

In the **Feature Engineering** process, this step involves breaking down the new_sentence column into simpler and more standardized formats. The process includes two sub-processes:

1. **Stemming :** uses PorterStemmer from NLTK to perform stemming on the words in the meaningful column, reducing them to their root forms and saving the output in a new column, stemmed_words. This process is aimed at normalizing words by removing suffixes, which helps reduce dimensionality and ensures consistency in text analysis, especially for machine learning tasks.
2. **Lemmatization** : uses WordNetLemmatizer to convert words in the meaningful column into their base forms, storing the results in a new column, lemmatized_words. The purpose is to normalize words for better consistency and improve the quality of text data for downstream NLP tasks.

B. Feature Engineering

In this part, multiple sub-processes were carried out to create additional features from the new_sentence column to enrich the dataset:

1. **Sentence Length & Word Count :** alculates the length of each sentence in the New_Sentence column (Length) and the total number of words in each sentence (Word_Count). These features provide quantitative metrics that help identify patterns in sentence structure or complexity, which can be useful for feature selection in machine learning models.
2. **Sentiment Analysis**
   Using the TextBlob library, the sentiment polarity of each sentence is calculated and stored in the sentiment_score column, with values ranging from -1 (negative) to 1 (positive). This feature helps capture the emotional tone of each sentence, providing critical insights for tasks like sentiment analysis or customer feedback classification.

C. Feature Evaluation

The aim of this stage is to evaluate the quality of the generated features and their importance for the target prediction:

1. **Label Encoding**
   uses `LabelEncoder` to convert text-based features (`words`, `meaningful`, `stemmed_words`, and `lemmatized_words`) into numerical values by first joining list elements into strings and then encoding them. The goal is to prepare these features for machine learning models, which require numerical inputs, enabling efficient processing and feature comparison during training.
2. **Correlation Analysis**
   Analyzed the correlation between features and the target variable using a heatmap. This step helps identify redundant or irrelevant features by assessing their linear relationships.
3. **Mutual Information**
   Applied Mutual Information analysis to measure the dependency between features and the target variable. Features with the highest dependency scores were prioritized for model training.

Git

https://github.com/tatashandharu15/Final_Project_NautixTech

Thank You

Rakamin
Academy