전남대학교
CHONNAM NATIONAL UNIVERSITY

Machine Learning

Disaster Tweets

Classification

Professor. Jin-Taek Seong

데이터사이언스 대학원
SAO MAVIN

Content

1. Project Overview
2. Project Performance Details
3. Problems and Solutions
4. Q&A

Project Overview

1.1 Introduction and Problem Definition

Inspiration from Real-world Issues, I had been searching for the competition relate with the **disaster** classification.

In Kaggle I joined the competition called “**Natural Language Processing with Disaster Tweets**”

Project Overview

1.1 Introduction and Problem Definition

Competition Challenge:

The competition objective is to classify tweets into '**disaster**' (1) and '**not-disaster**'(0) categories by using the given data included text with keywords and additional metadata from **Tweet**.

id	keyword	location	text	target
1	ablaze	Paranaque City	Ablaze for you Lord :D	0
2	ablaze	Edmonton, Alberta - Treaty 6	How the West was burned: Thousands of wildfires ablaze in #..	1
3	ablaze	Concord, CA	@Navista7 Steve these fires out here are something else! California is a tinderbox - and this clown ...	1

Project Overview

1.2 Introduction to the proposed method to solve the problem

So, we chooses and built the baseline model by using **TF-IDF (Term Frequency-Inverse Document Frequency)** with **Logistic Regression**, **Naïve Bayes** and **SVM.**

By using these machine learning models the performance and accuracy is work significantly on each model. But our task is to improve these baseline model to make it perform more better with a better accuracy score.

Project Overview

1.2 Introduction to the proposed method to solve the problem

To enhance machine learning models such as **Logistic Regression**, **Naïve Bayes**, **and SVM** baseline model. I'll do this using three techniques:

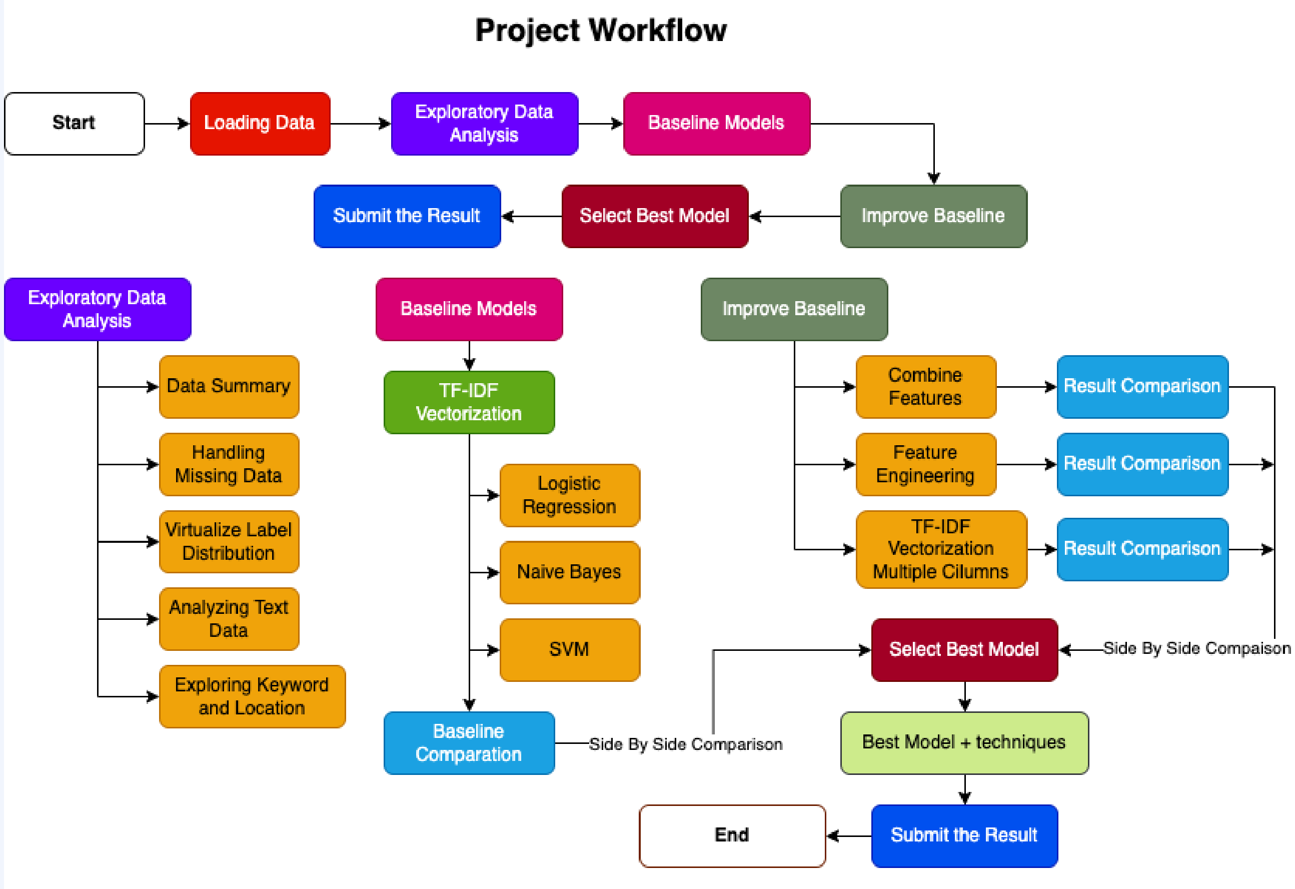
Combine Text and Keyword Features : Merge information from **'text'** and **'keyword'** columns to capture complementary patterns.

Feature Engineering : Create or **modify features** to enhance the model's understanding and predictive power**.**

TF-IDF Vectorization for Multiple Columns : Convert **text** and **categorical features** into numerical vectors for improved model comprehension.

2. Project Performance Details

2.1 How to carry out the project

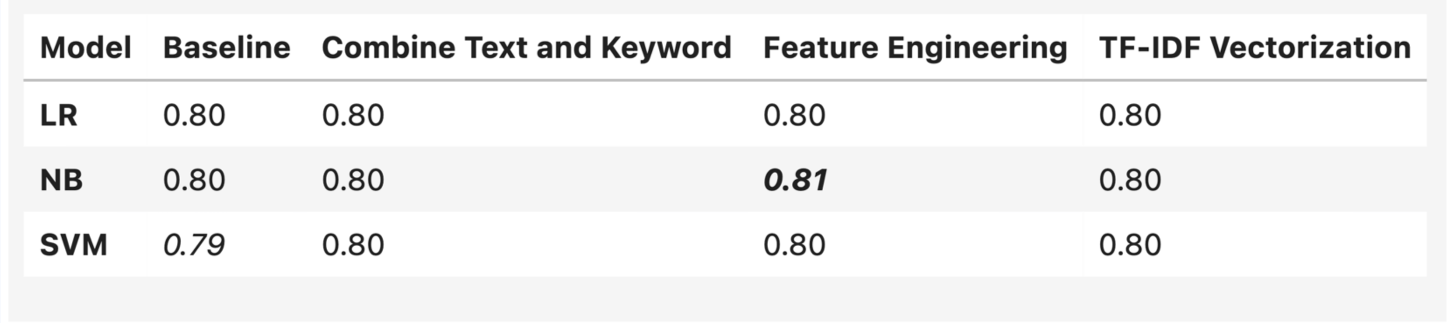


2. Project Performance Details

2.2 Project performance results

The results obtained through the proposed method are summarized. The tables below are the comparison side by side between all the technique we use to improve the baseline model.

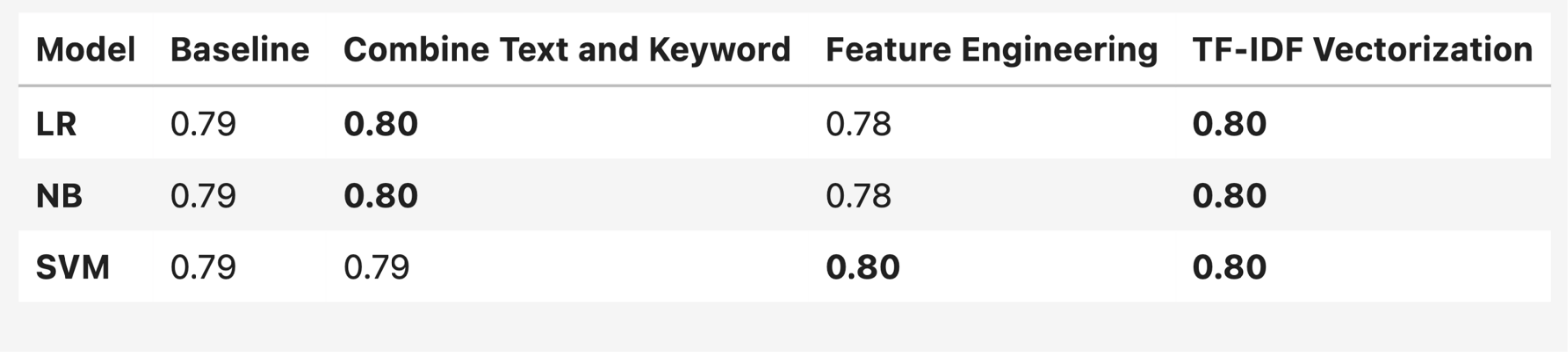
2.2.1 Accuracy Improvement Table



Model	Baseline	Combine Text and Keyword	Feature Engineering	TF-IDF Vectorization
LR	0.80	0.80	0.80	0.80
NB	0.80	0.80	***0.81***	0.80
SVM	*0.79*	0.80	0.80	0.80

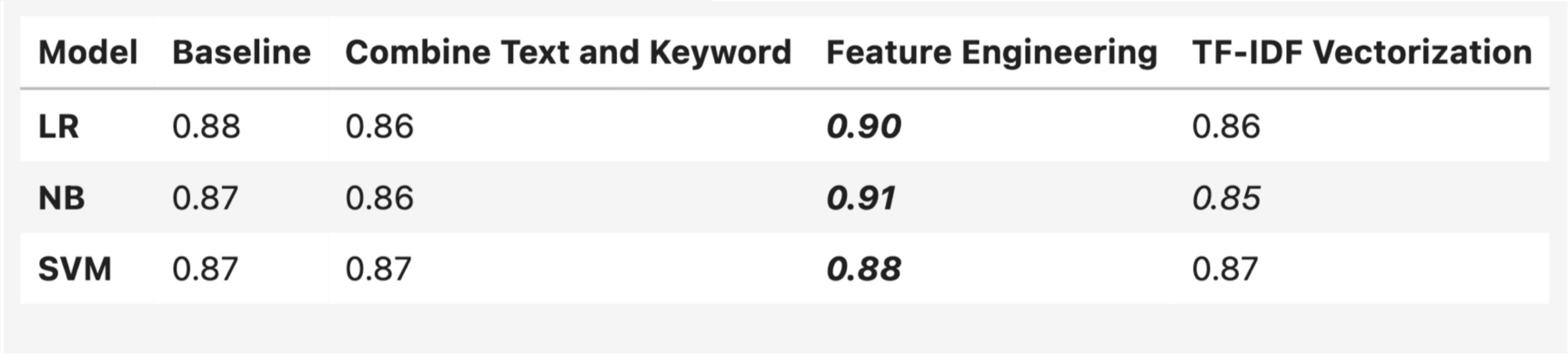
2. Project Performance Details

2.2.2 Precision Improvement Table



Model	Baseline	Combine Text and Keyword	Feature Engineering	TF-IDF Vectorization
LR	0.79	**0.80**	0.78	**0.80**
NB	0.79	**0.80**	0.78	**0.80**
SVM	0.79	0.79	**0.80**	**0.80**

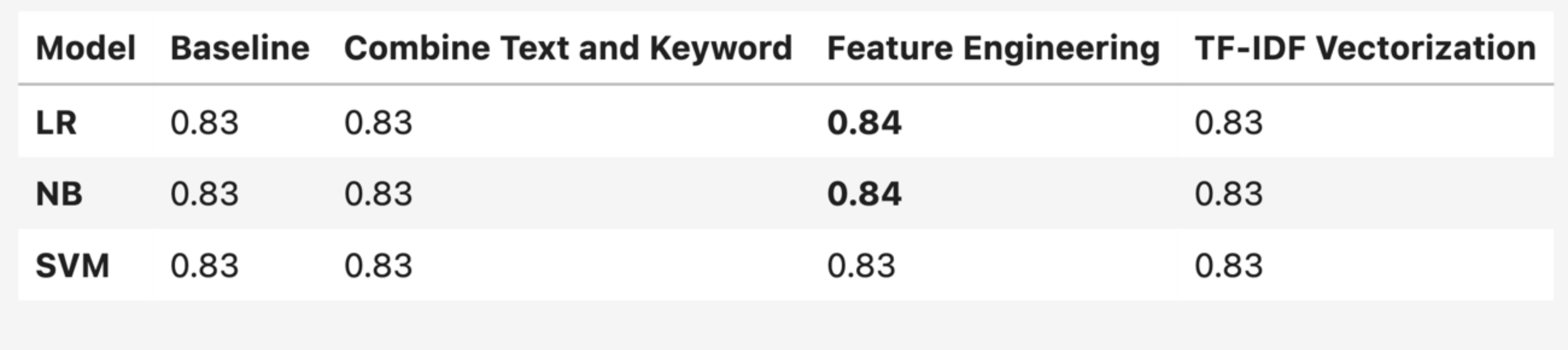
2.2.3 Recall Improvement Table



Model	Baseline	Combine Text and Keyword	Feature Engineering	TF-IDF Vectorization
LR	0.88	0.86	***0.90***	0.86
NB	0.87	0.86	***0.91***	*0.85*
SVM	0.87	0.87	***0.88***	0.87

2. Project Performance Details

2.2.4 F1 Improvement Table



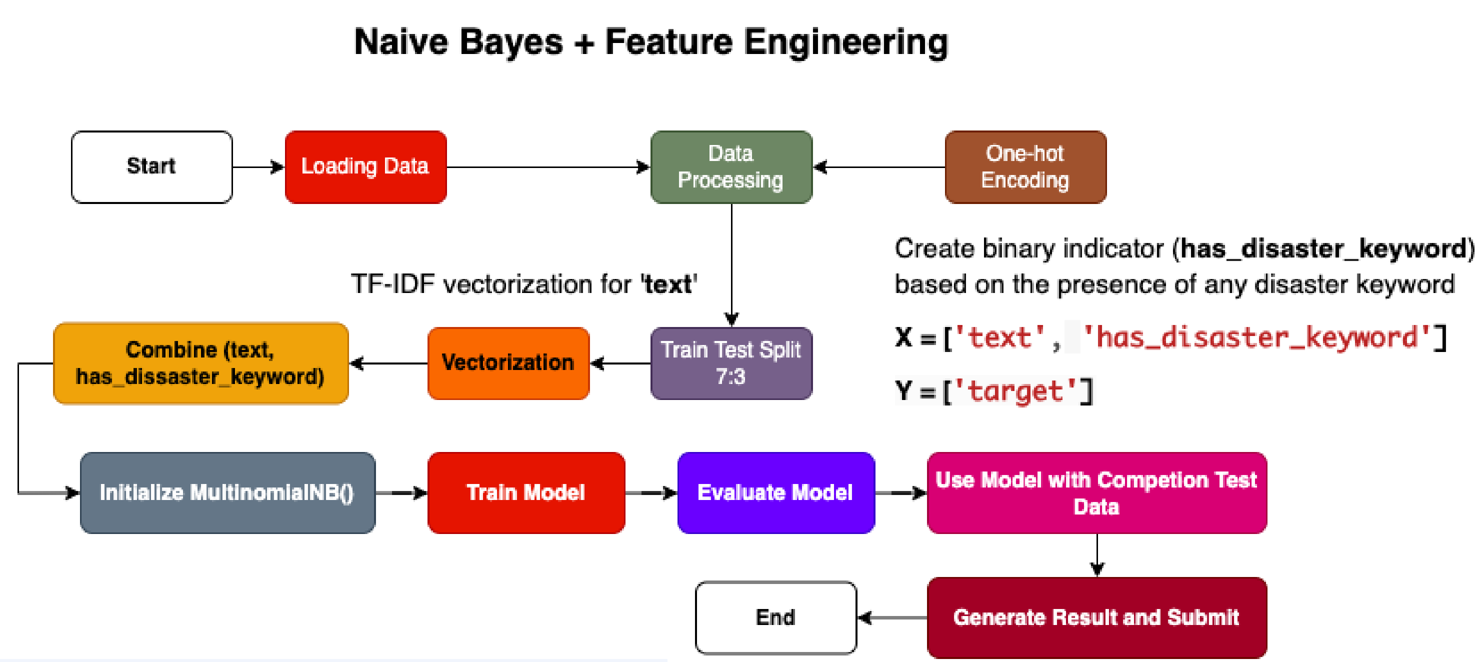
Model	Baseline	Combine Text and Keyword	Feature Engineering	TF-IDF Vectorization
LR	0.83	0.83	**0.84**	0.83
NB	0.83	0.83	**0.84**	0.83
SVM	0.83	0.83	0.83	0.83

2.3 Result Conclusion

Base on the results of performance(Accuracy, Precision, Recall, and F1) we choose **Naive Bayes (NB) with Feature Engineering technique** for the competition. Because, NB showed improvement in accuracy, precision, recall, and F1 score. Feature Engineering enhanced its overall performance.

2. Project Performance Details

2.4 Naïve Bayes + Feature Engineering Workflow


Naive Bayes + Feature Engineering
Start
Loading Data
Data Processing
One-hot Encoding
Create binary indicator (has_disaster_keyword) based on the presence of any disaster keyword
X = ['text', 'has_disaster_keyword']
Y = ['target']
TF-IDF vectorization for 'text'
Train Test Split 7:3
Vectorization
Combine (text, has_dissaster_keyword)
Initialize MultinomialNB()
Train Model
Evaluate Model
Use Model with Competion Test Data
Generate Result and Submit
End

2. Project Performance Details

2.3 Main code description

Method name	Input/Output
preprocess_text	• **Input:** The input text data that requires cleaning and preprocessing. • **Output**: A cleaned version of the input text data after performing a series of cleaning operations.
get_top_words	• **Text:** The input text that the function will analyze to extract the top words. • **Number of Words (n):** A numerical parameter that specifies the desired count of top words to be returned. • **Output:** A list or collection of the top N words based on the provided text and the specified number of words (n).
eval_matrix	• **eval_labels:** The actual labels used for evaluation. • **predictions:** The predicted labels generated by the model. • **Output:** A visual representation of the confusion matrix and generates a report summarizing the model's performance.

2. Project Performance Details

2.3 Main code description

Method name	Input/Output
fe_data_preparation	• **dataframe:** The input dataframe containing text data. • **Output:** a modified dataframe that includes a new binary feature, "**has_disaster_keyword**," based on the presence of any disaster-related keywords in the text.
plot_top_words	• **ax:** Subplot axes where the histogram will be plotted. • **top_words:** A list of the top N words to be visualized on the histogram • **title:** The title of the plot. • **Output:** prepares the subplot **axes (ax)** for plotting the histogram of the top **N** words. It doesn't explicitly return anything but sets up the environment for subsequent plotting.

3. Problems and Solutions

Related to Project Implementation

3.1 Problem identification

Project Planning Adjustments:

- Adjustments on original plan outline in my plan report.

Technical Challenges:

- Faced numerous errors in the implementation phase.
- Required substantial time and effort to troubleshoot and resolve issues.

3. Problems and Solutions

Related to Project Implementation

3.2 Solution

- Revised the plan to focus on enhancing the baseline model through exploration of techniques.
- Utilize shared code and discussion sections in the competition for **problem-solving**.
- Leverage community support when encountering challenging issues.
- Seek assistance when stuck on ideas or in need of clarification on specific topics. Tap into **ChatGPT** for insights and guidance.

3. Problems and Solutions

Related to Project Implementation

3.3 Project Performance Review

Enhanced Problem-Solving Skills:

- The hands-on experience contributed to the enhancement of problem-solving skills.
- Dealt with real-world challenges to develop a practical understanding.

Application of Knowledge:

- The competition served as a valuable opportunity to apply acquired knowledge. Applied concepts learned throughout the semester in a real-world context.

4. References

- [1] Seong Jin-taek, “Project Guide” Chonnam National University Graduate School of Data Science, 2023.

Thank You

Q&A