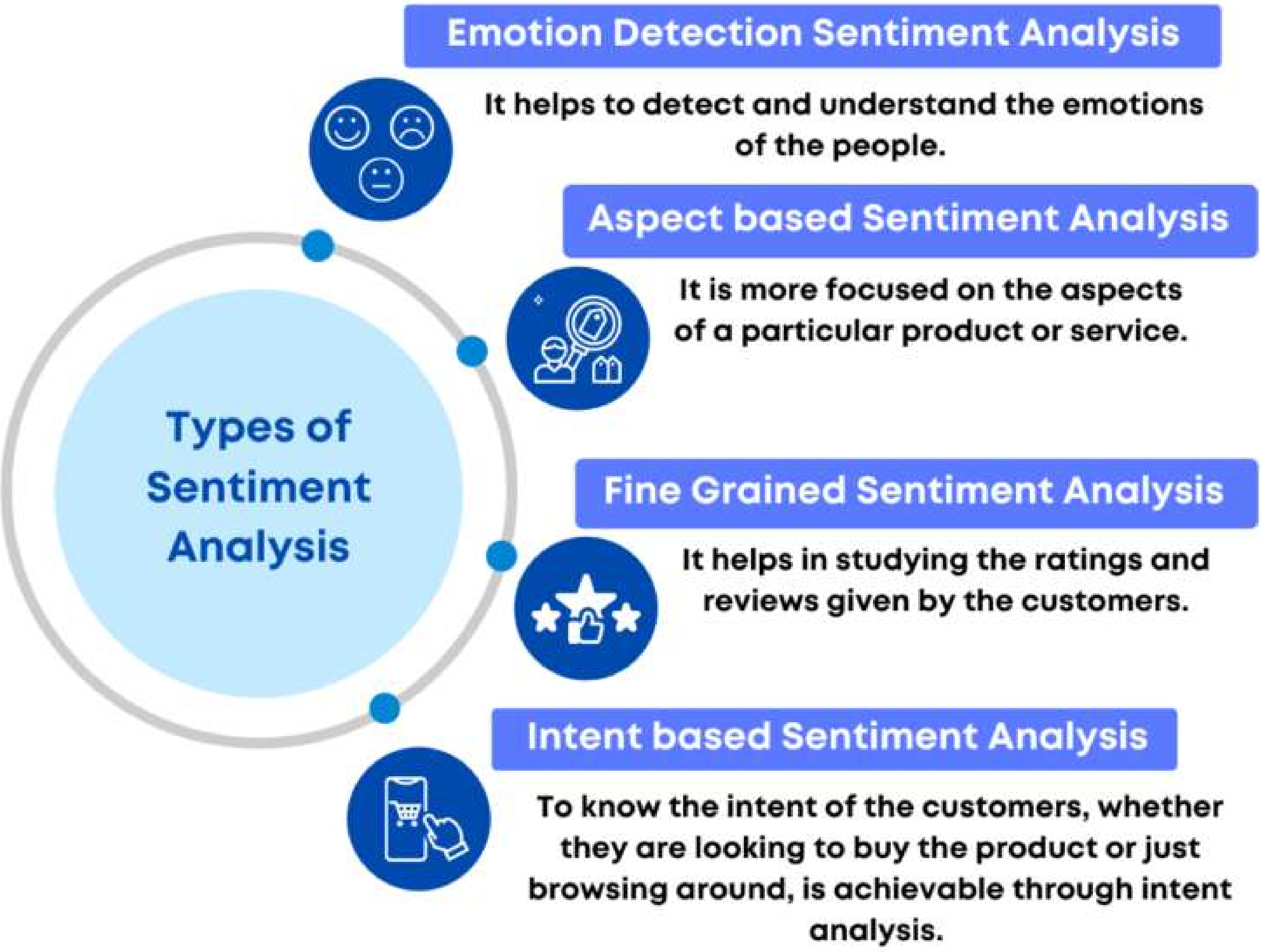
SENTIMENT ANALYSIS



Types of Sentiment Analysis

Emotion Detection Sentiment Analysis

It helps to detect and understand the emotions of the people.

Aspect based Sentiment Analysis

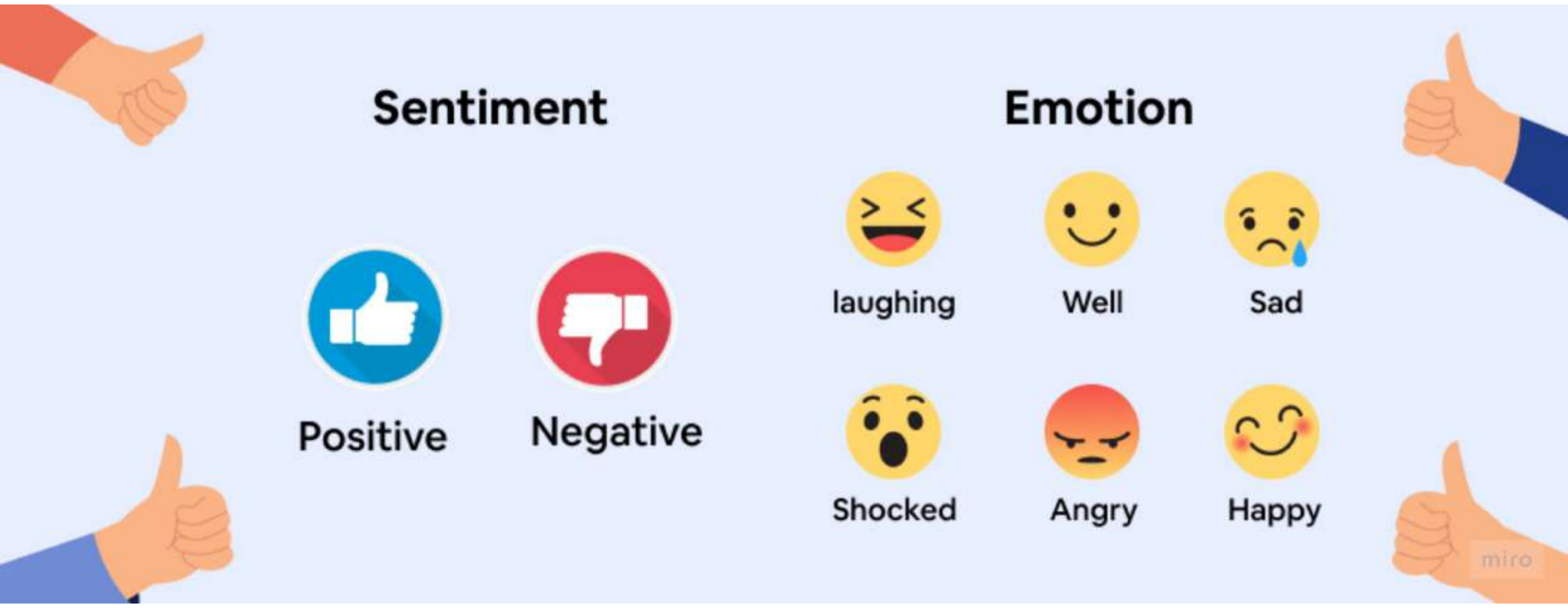
It is more focused on the aspects of a particular product or service.

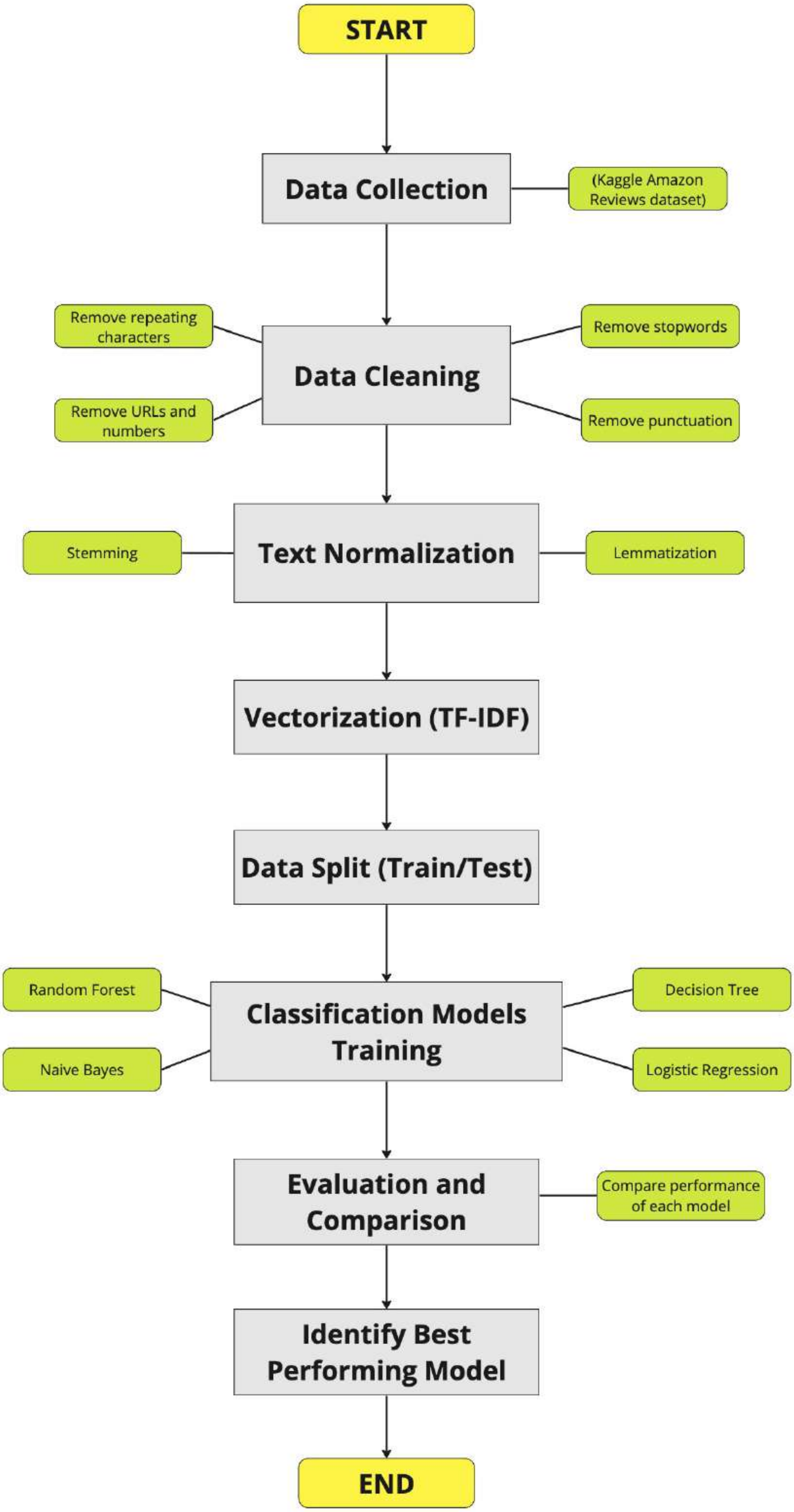
Fine Grained Sentiment Analysis

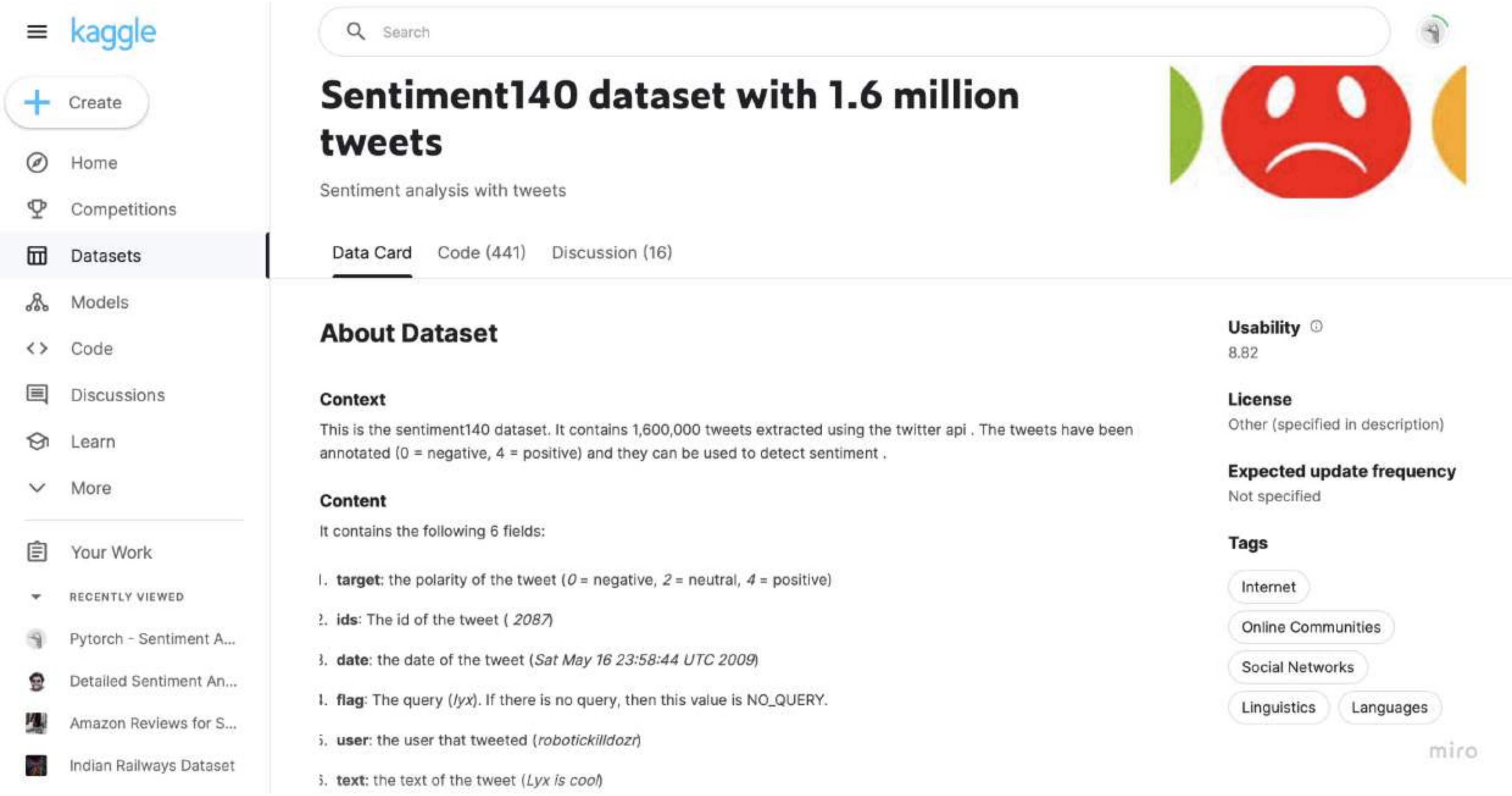
It helps in studying the ratings and reviews given by the customers.

Intent based Sentiment Analysis

To know the intent of the customers, whether they are looking to buy the product or just browsing around, is achievable through intent analysis.


Sentiment
Positive
Negative
Emotion
laughing
Well
Sad
Shocked
Angry
Happy


START
Data Collection
(Kaggle Amazon Reviews dataset)
Remove repeating characters
Remove stopwords
Data Cleaning
Remove URLs and numbers
Remove punctuation
Stemming
Text Normalization
Lemmatization
Vectorization (TF-IDF)
Data Split (Train/Test)
Random Forest
Decision Tree
Classification Models Training
Naive Bayes
Logistic Regression
Evaluation and Comparison
Compare performance of each model
Identify Best Performing Model
END
miro



Sentiment140 dataset with 1.6 million tweets

Sentiment analysis with tweets

About Dataset

Context

This is the sentiment140 dataset. It contains 1,600,000 tweets extracted using the twitter api . The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment .

Content

It contains the following 6 fields:

1. **target**: the polarity of the tweet (*0* = negative, *2* = neutral, *4* = positive)
2. **ids**: The id of the tweet (*2087*)
3. **date**: the date of the tweet (*Sat May 16 23:58:44 UTC 2009*)
4. **flag**: The query (*lyx*). If there is no query, then this value is NO_QUERY.
5. **user**: the user that tweeted (*robotickilldozr*)
6. **text**: the text of the tweet (*Lyx is cool*)

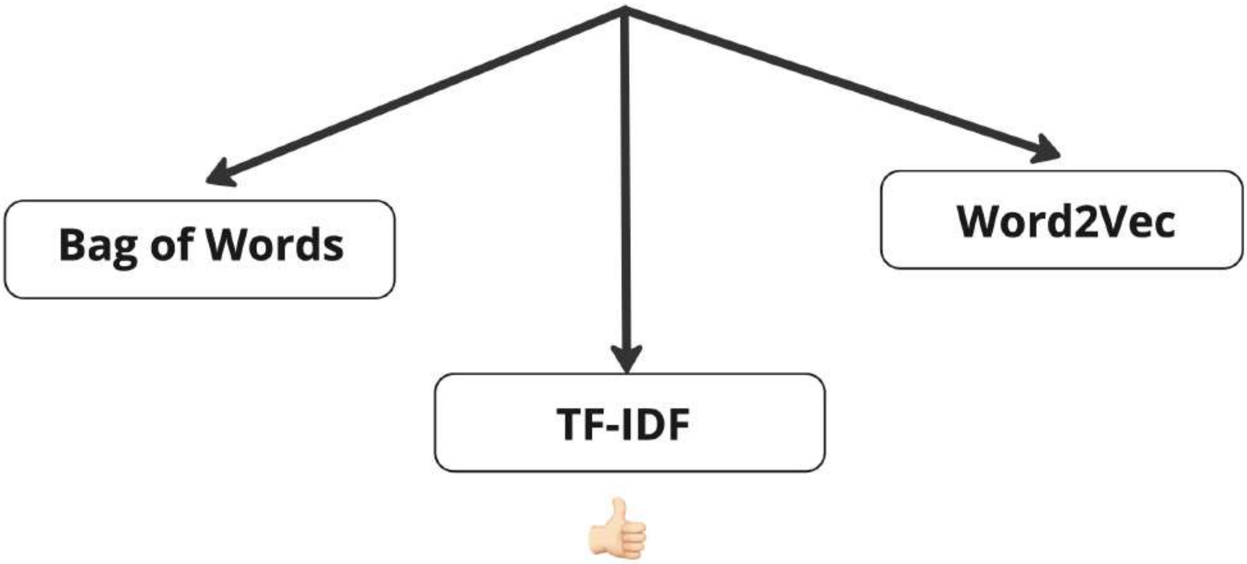
Usability
8.82

License
Other (specified in description)

Expected update frequency
Not specified

Tags
Internet
Online Communities
Social Networks
Linguistics
Languages

Vectorization

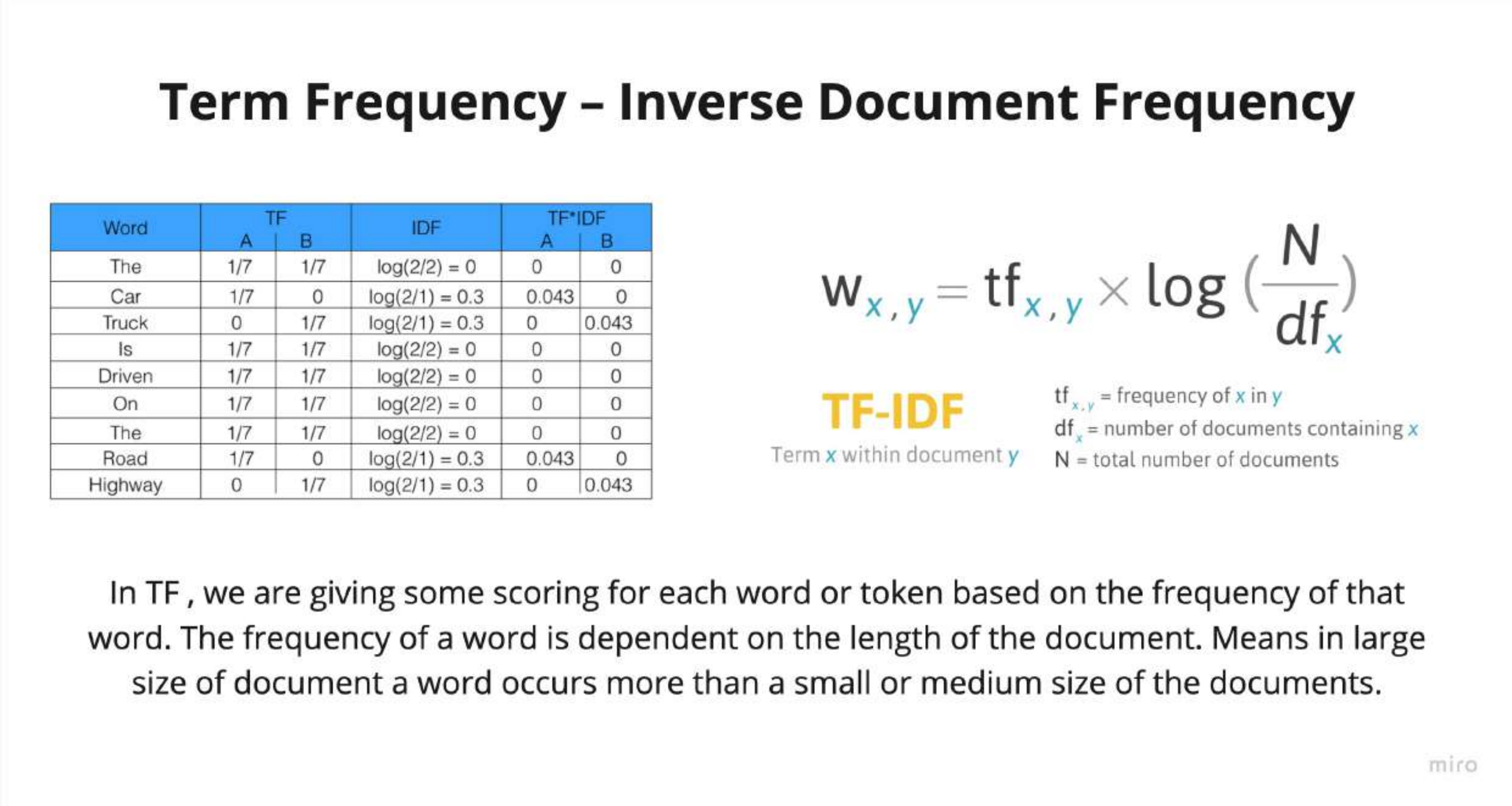


Bag of Words

TF-IDF

👍

Word2Vec



Term Frequency – Inverse Document Frequency

Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	log(2/2) = 0	0	0
Car	1/7	0	log(2/1) = 0.3	0.043	0
Truck	0	1/7	log(2/1) = 0.3	0	0.043
Is	1/7	1/7	log(2/2) = 0	0	0
Driven	1/7	1/7	log(2/2) = 0	0	0
On	1/7	1/7	log(2/2) = 0	0	0
The	1/7	1/7	log(2/2) = 0	0	0
Road	1/7	0	log(2/1) = 0.3	0.043	0
Highway	0	1/7	log(2/1) = 0.3	0	0.043

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

In TF , we are giving some scoring for each word or token based on the frequency of that word. The frequency of a word is dependent on the length of the document. Means in large size of document a word occurs more than a small or medium size of the documents.

Why Classification Model ?

- Classification models learn patterns in text data indicative of sentiment.
- They capture linguistic cues, contextual information, and sentiment-related features.
- Classification models generalize well to unseen data, making them applicable to different datasets and real-world scenarios.
- They offer flexibility in model selection and feature engineering.
- Algorithms like Naive Bayes, Random Forest, and Logistic Regression have different strengths for text classification tasks.
- Techniques like TF-IDF vectorization and word embeddings can enhance model performance.

Other Techniques

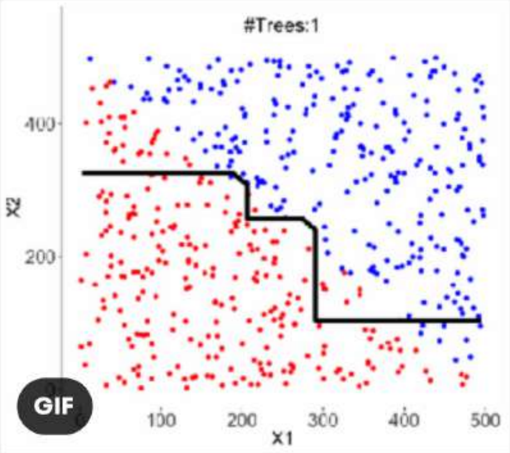
1. **Rule-based approaches**: These methods use predefined linguistic rules or sentiment lexicons to determine sentiment. While interpretable, they may struggle to capture complex relationships in the data.

2. **Deep learning models**: Models like RNNs, CNNs, and Transformer-based architectures (e.g., BERT) have shown promise in sentiment analysis. They automatically learn hierarchical representations and capture intricate dependencies, but require substantial labeled data and computational resources for training.

3. **Ensemble methods**: Techniques like Bagging and Boosting combine multiple models or predictions to enhance performance. They can be applied to sentiment analysis by aggregating predictions from multiple classifiers, improving accuracy.

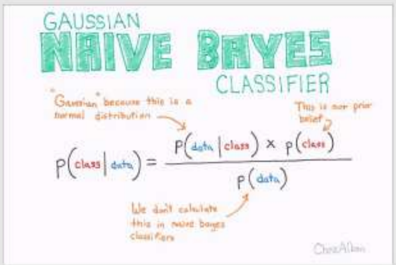
4. **Hybrid approaches**: These methods combine rule-based, machine learning, or deep learning techniques to leverage their respective strengths. By blending interpretability and predictive power, they aim to improve sentiment analysis results.

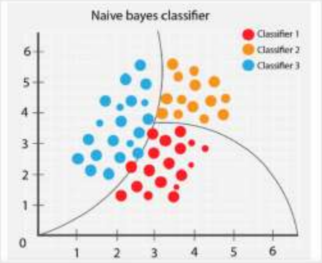
Random Forest



- Random Forest: Ensemble learning algorithm for classification and regression.
- Combines decision trees for improved accuracy.
- Uses random subsets of features and samples for each tree.
- Final prediction obtained through averaging or voting.
- Handles high-dimensional data, mitigates overfitting.
- Provides feature importance rankings.

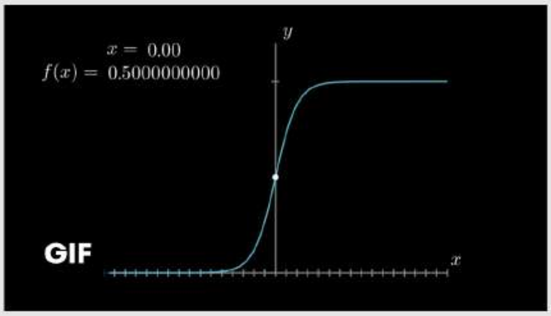
Naive Bayes


GAUSSIAN
NAIVE BAYES
CLASSIFIER
"Gaussian" because this is a normal distribution
This is our prior belief
P(class | data) = P(data | class) × p(class) / P(data)
We don't calculate this in naive bayes classifiers
ChrisAlbon

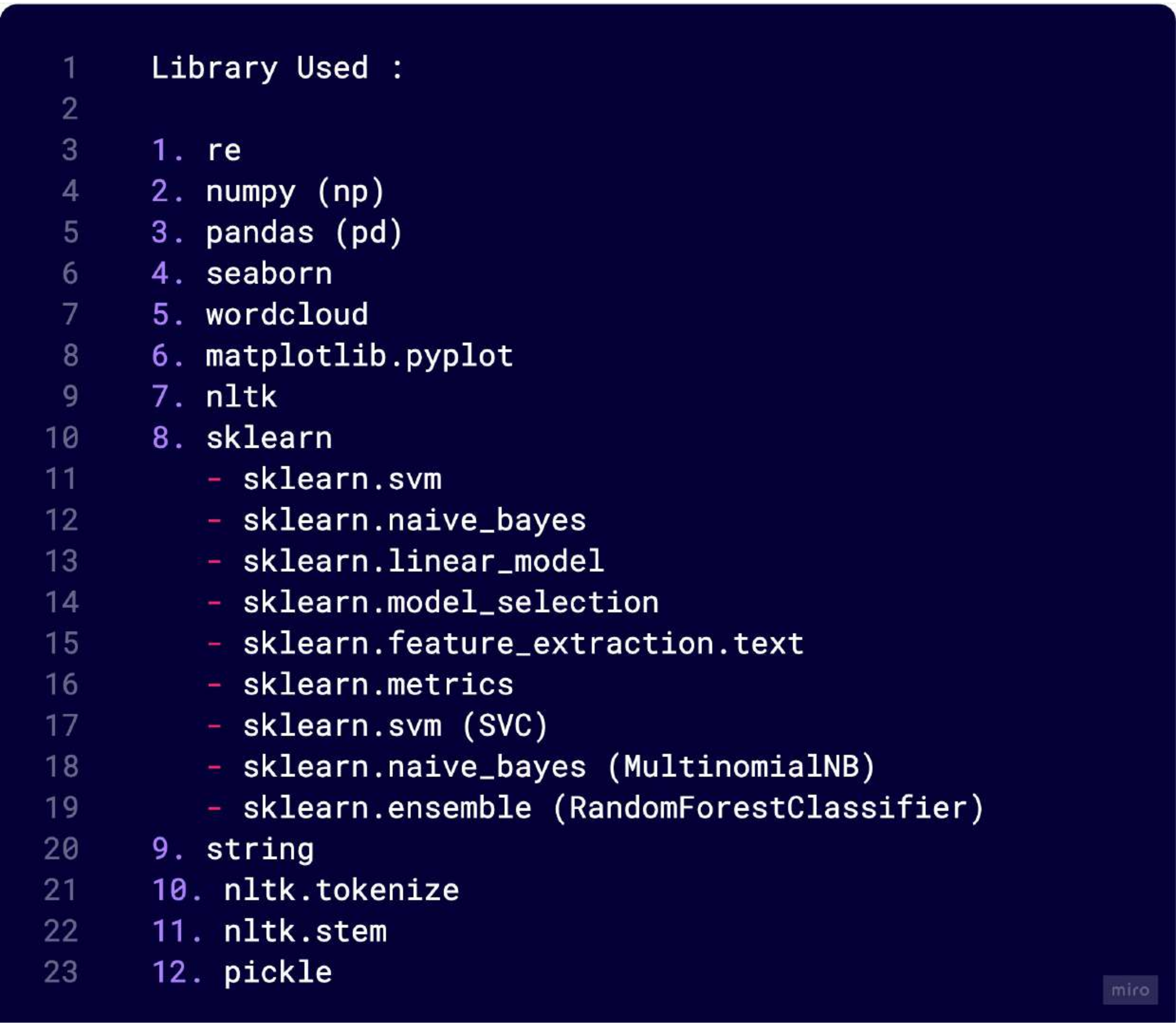


- Naive Bayes is a simple and efficient classification algorithm.
- It is based on Bayes' theorem and assumes independence between features.
- Naive Bayes is particularly effective for text classification tasks.
- It calculates the probability of each class given the input features.
- Naive Bayes handles high-dimensional data and is computationally efficient.
- Despite its "naive" assumption, Naive Bayes often performs well in practice.

Logistic Regression

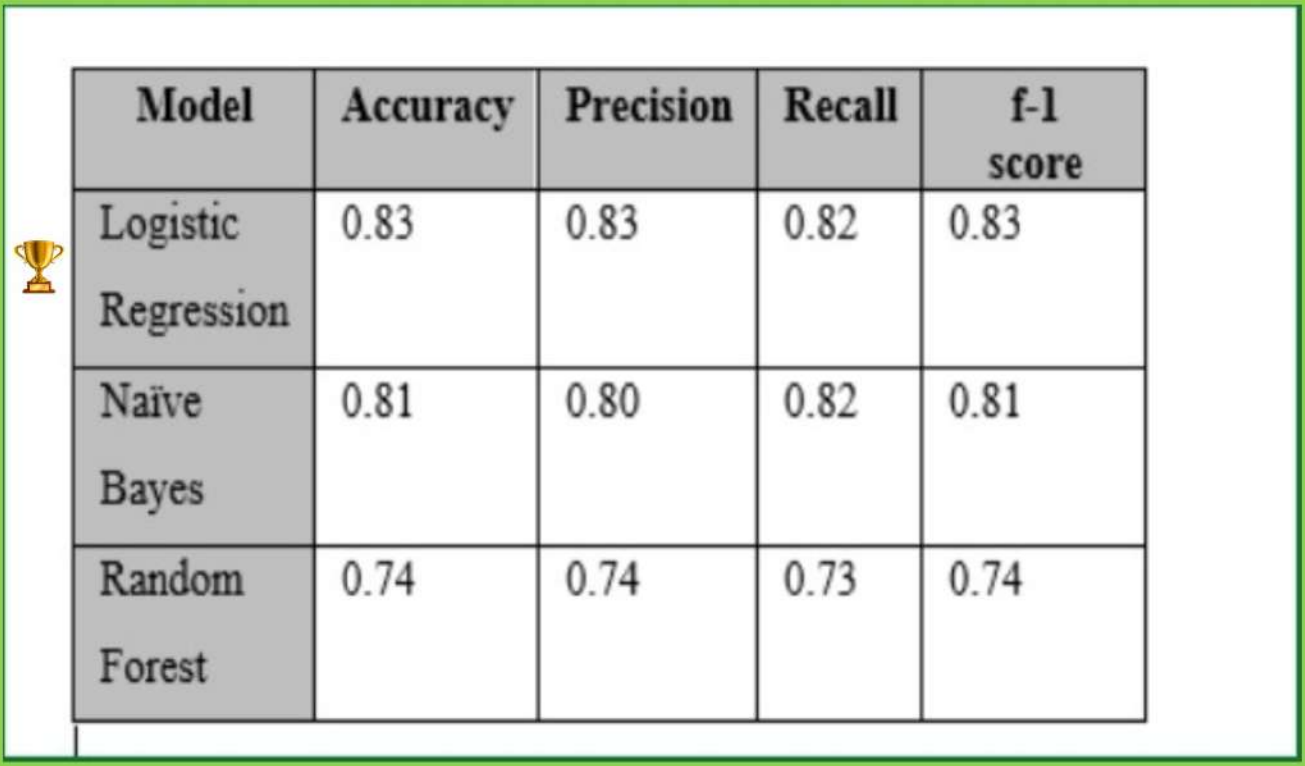


- Logistic Regression is a widely used statistical model for binary classification.
- It estimates the probability of an instance belonging to a certain class.
- It uses a logistic function (sigmoid) to model the relationship between input variables and the predicted probability.
- Logistic Regression can handle both numerical and categorical input features.
- It is interpretable, allowing for easy understanding of the impact of input variables.
- Logistic Regression is computationally efficient and can handle large datasets.
- It is commonly used in various domains, including sentiment analysis, disease prediction, and customer churn analysis.
- Logistic Regression is a powerful tool for binary classification tasks, providing insights and predictive capabilities.
- Overall, Logistic Regression is a reliable and interpretable method for making binary predictions based on input features.



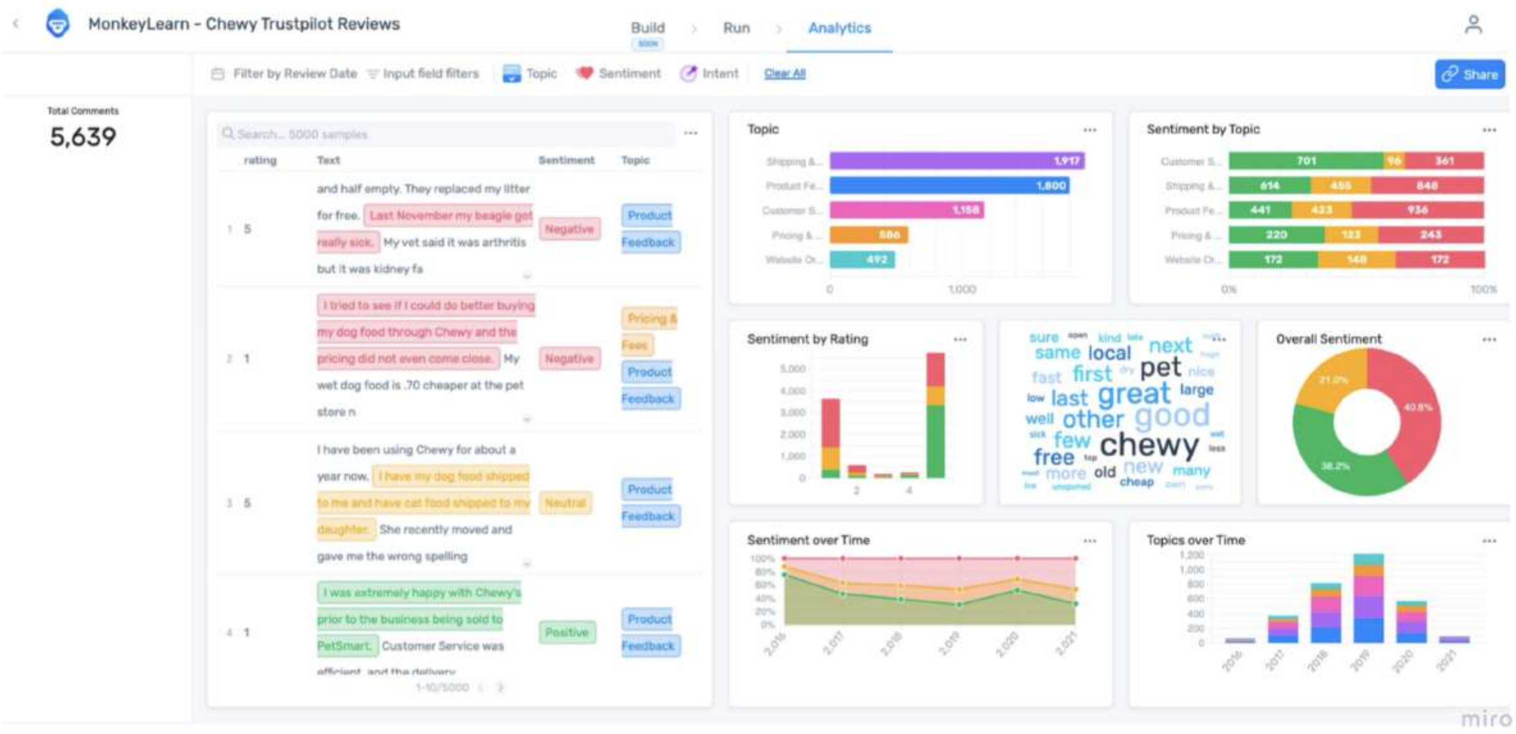
```
Library Used :

1. re
2. numpy (np)
3. pandas (pd)
4. seaborn
5. wordcloud
6. matplotlib.pyplot
7. nltk
8. sklearn
   - sklearn.svm
   - sklearn.naive_bayes
   - sklearn.linear_model
   - sklearn.model_selection
   - sklearn.feature_extraction.text
   - sklearn.metrics
   - sklearn.svm (SVC)
   - sklearn.naive_bayes (MultinomialNB)
   - sklearn.ensemble (RandomForestClassifier)
9. string
10. nltk.tokenize
11. nltk.stem
12. pickle
```

Research Conclusion



Model	Accuracy	Precision	Recall	f-1 score
🏆 Logistic Regression	0.83	0.83	0.82	0.83
Naïve Bayes	0.81	0.80	0.82	0.81
Random Forest	0.74	0.74	0.73	0.74

Application of Text based Sentiment Analysis in Real Word



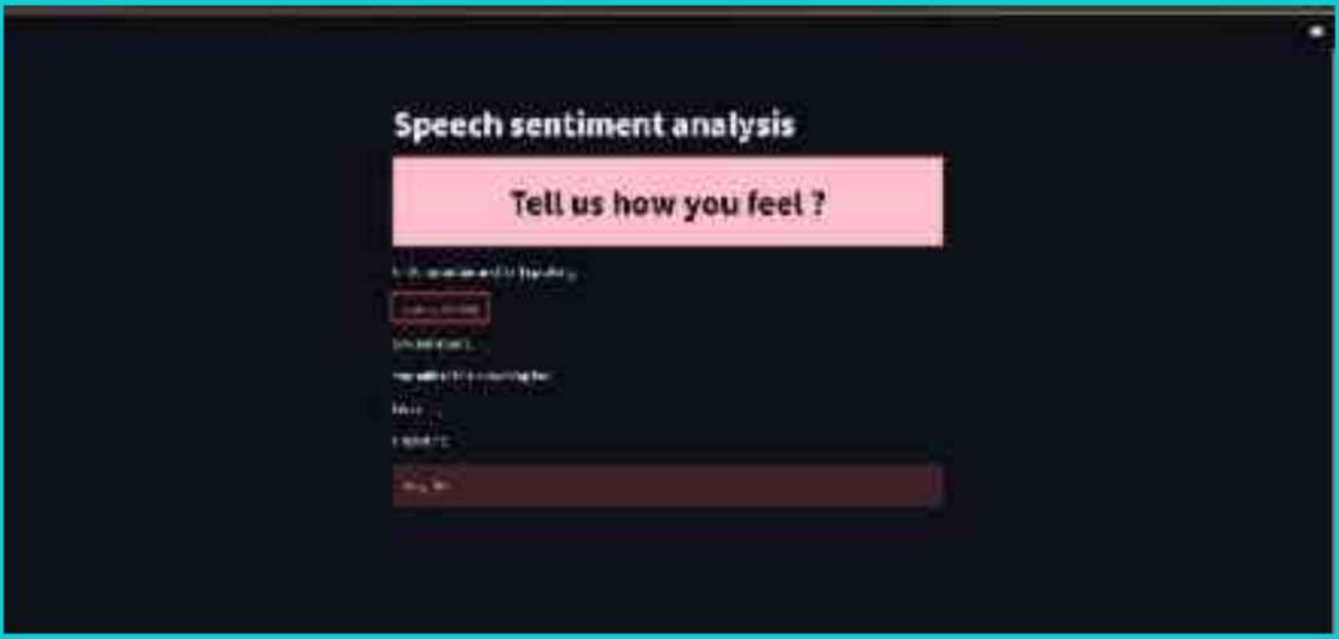
	rating	Text	Sentiment	Topic
1	5	and half empty. They replaced my litter for free. Last November my beagle got really sick. My vet said it was arthritis but it was kidney fa	Negative	Product Feedback
2	1	I tried to see if I could do better buying my dog food through Chewy and the pricing did not even come close. My wet dog food is .70 cheaper at the pet store n	Negative	Pricing & Fees Product Feedback
3	5	I have been using Chewy for about a year now. I have my dog food shipped to me and have cat food shipped to my daughter. She recently moved and gave me the wrong spelling	Neutral	Product Feedback
4	1	I was extremely happy with Chewy's prior to the business being sold to PetSmart. Customer Service was efficient and the delivery	Positive	Product Feedback

1-10/5000

Project

Deployed ML based Application depicting the use case of Sentiment Analysis.

Here we start with **Audio Modality** Sentiment Analysis



Speech sentiment analysis

Tell us how you feel ?