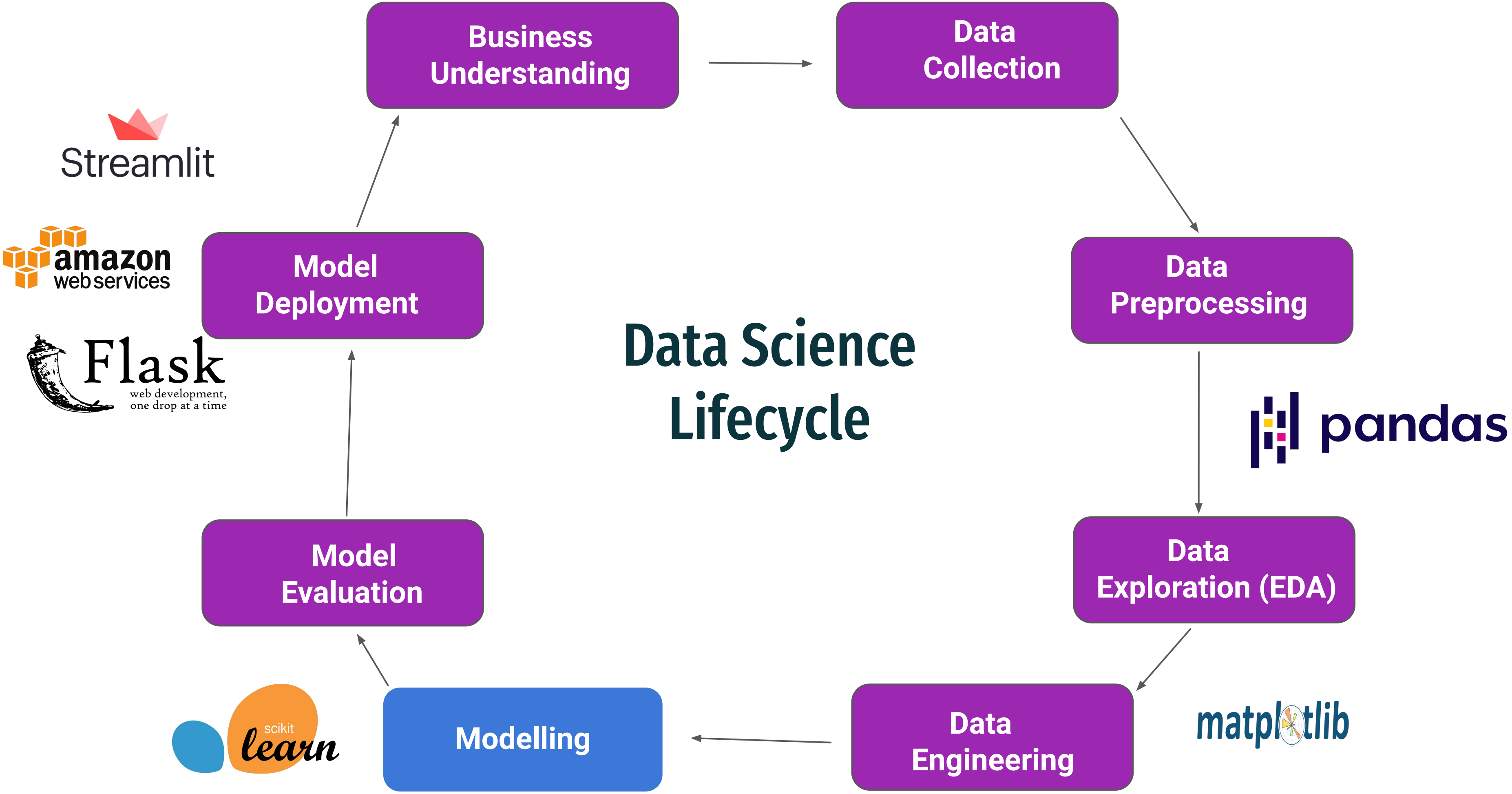
2022 CLIMATE CHANGE BELIEF ANALYSIS

Data Science Lifecycle



Business Understanding

Data Collection

Data Preprocessing

Data Exploration (EDA)

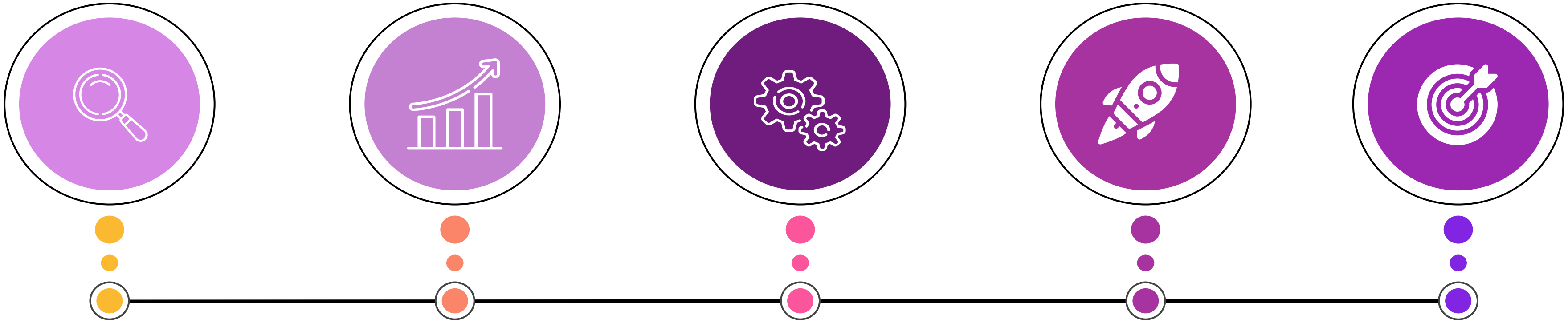
Data Engineering

Modelling

Model Evaluation

Model Deployment

Natural Language Processing Lifecycle



Data Exploration (EDA)

Data Visualization
Feature Analysis

Data Preprocessing

Data Cleaning

Feature Text Engineering

Normalizing
Vectorizing
Scaling
Imbalanced Data

Hyperparameter Tuning

GridSearchCV
Model Evaluation

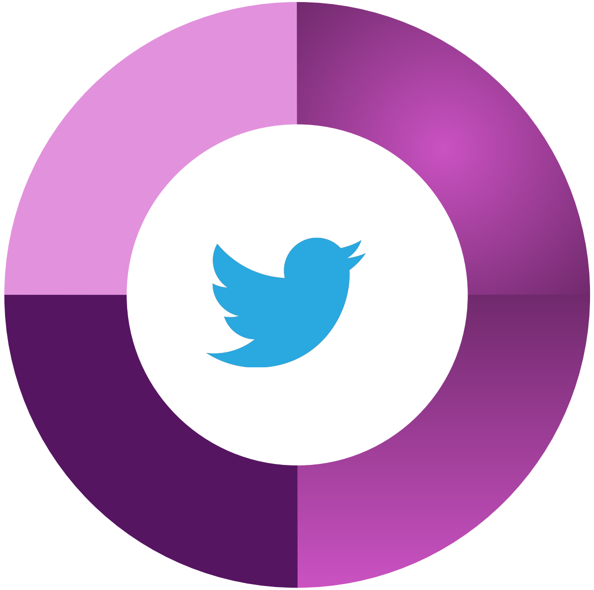
Build Model

Visualizing
Model Performance

Understanding Problem Statement

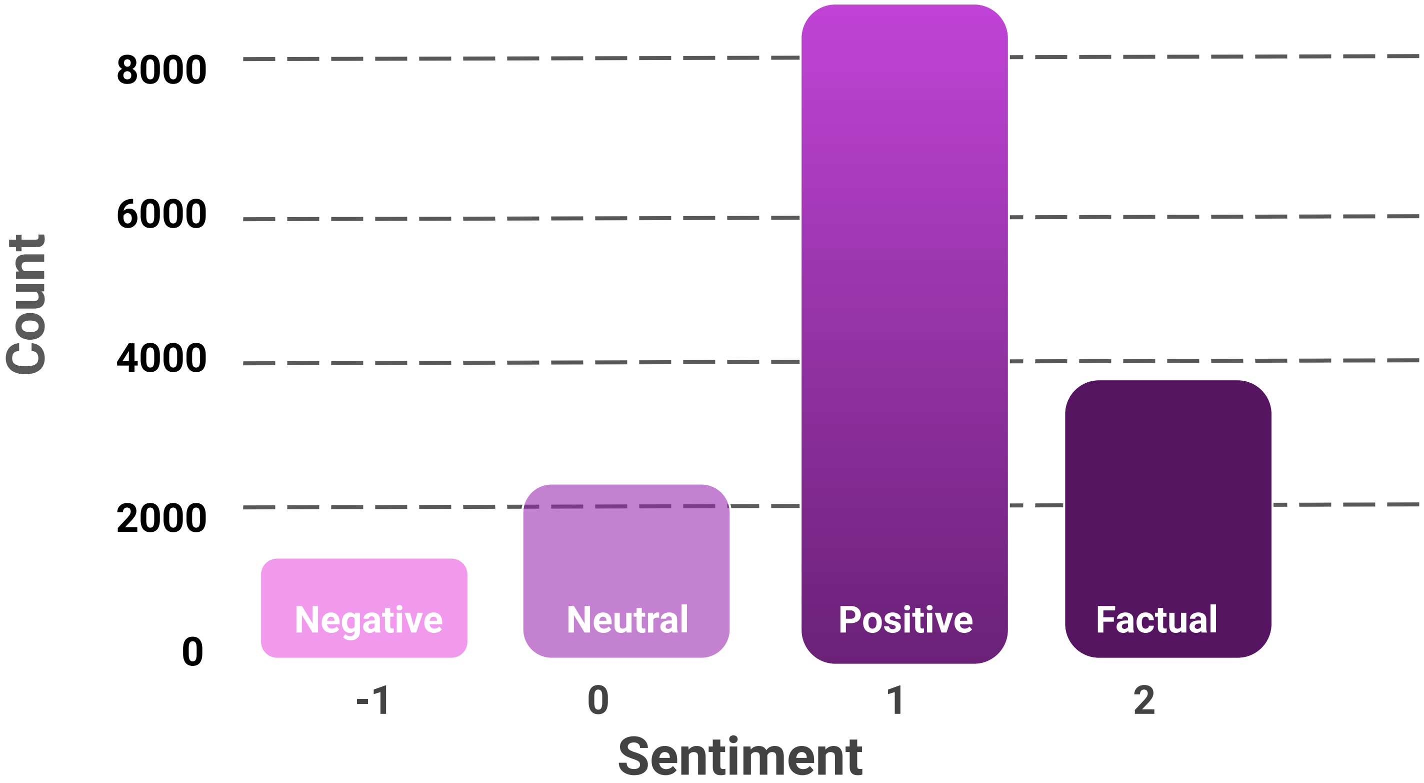
- **Translate Language**
  Machine Learning Techniques to translate language

- **Climate Change**
  All humans have the fundamental rights to live in a sustainable environment.

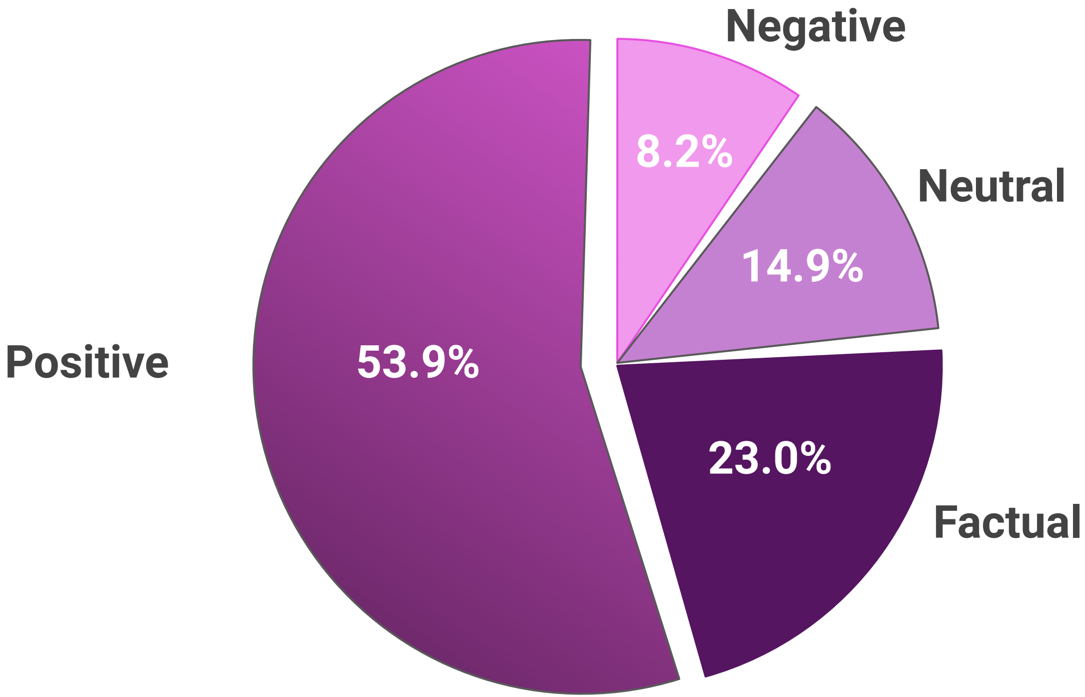


- **Sentimental Insights**
  Drawing insights from what people have twitted about climate change

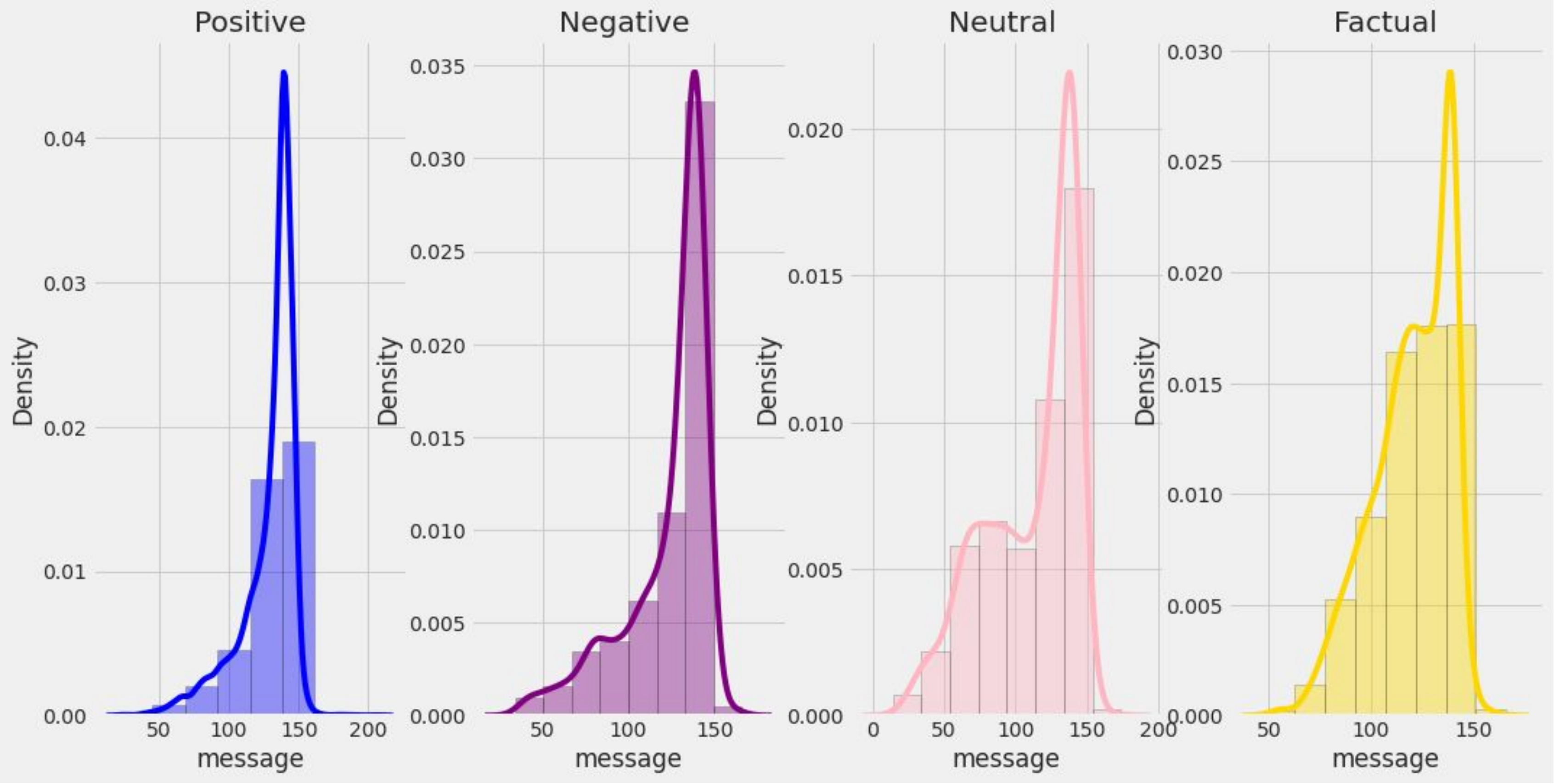
Number of Messages Per Sentiment



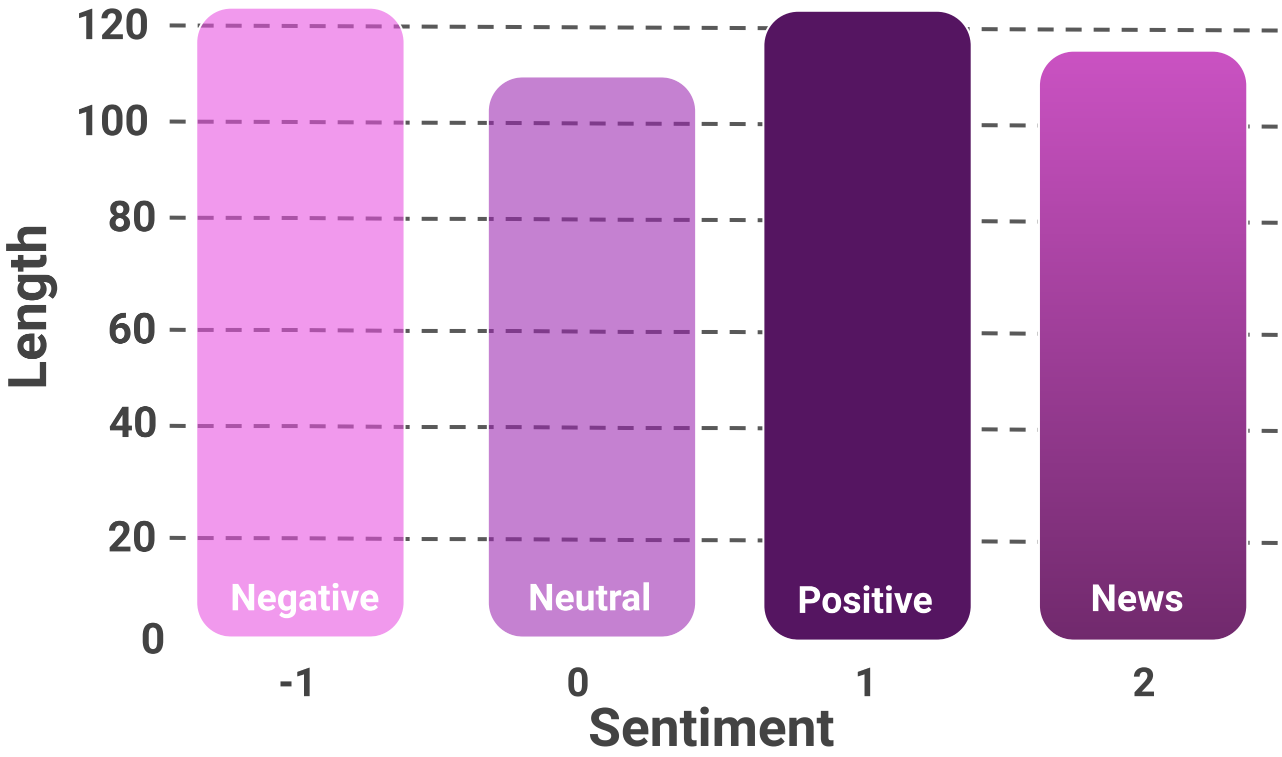
Percentage of Messages per Sentiment



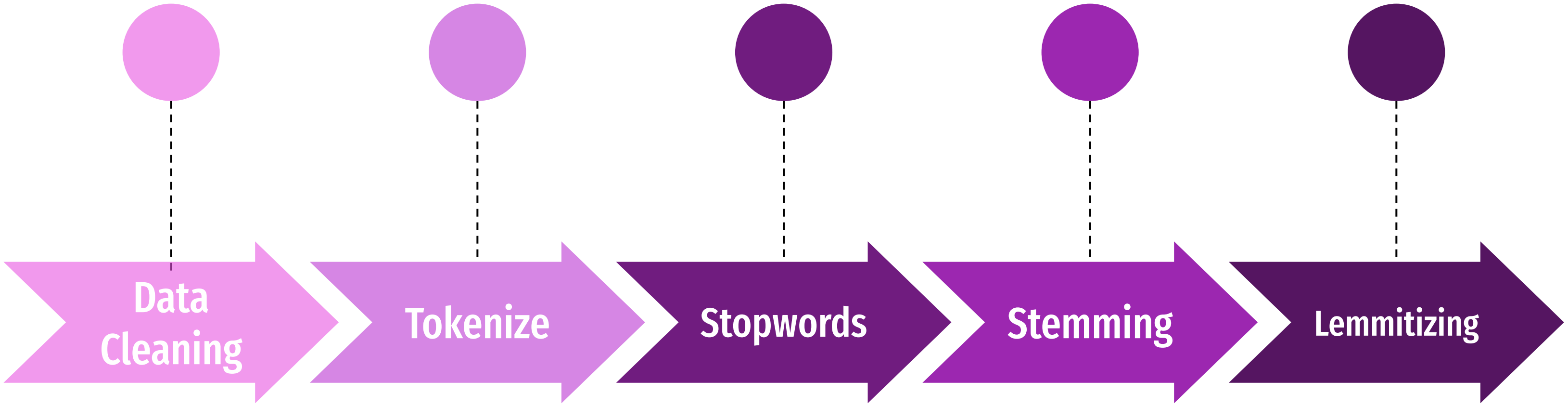
Distribution of Length per Label



Average Length of Messages by Sentiment



Data Preprocessing



Data Cleaning

Tokenize

Stopwords

Stemming

Lemmitizing

Remove Noise

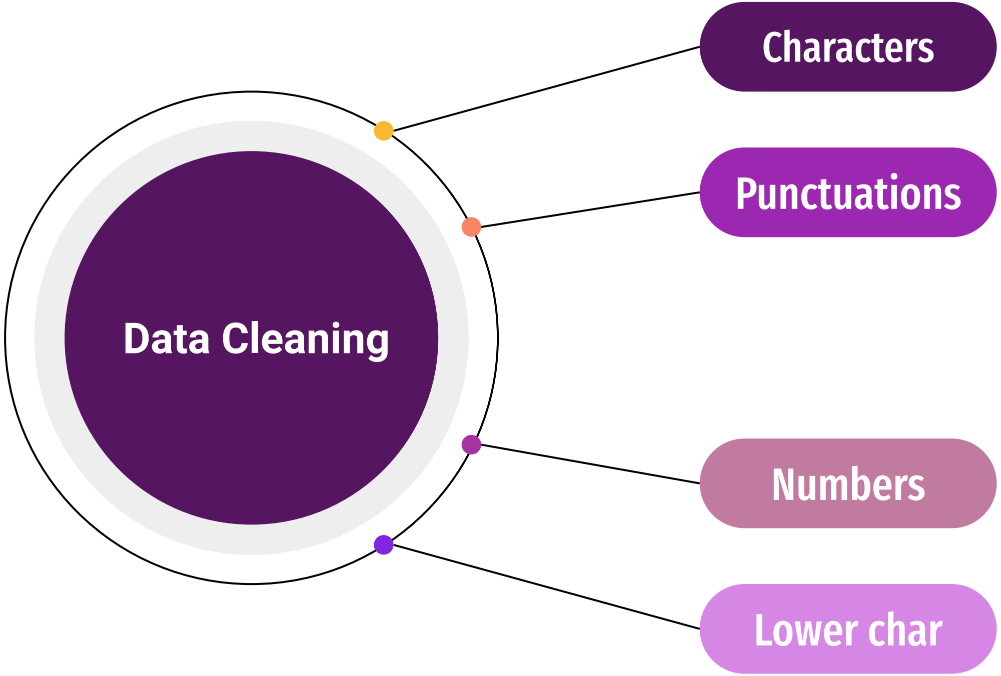
Tokens of strings

English Stopwords

Roor meaning of a word

Returning the dictionary meaning of a word

Data Preprocessing



Data Cleaning

Characters

Hashtags
URLs
Retweets
Mentions
Emojis

Punctuations

Remove [!*&%_-?\/.,<,>]

Numbers

Remove Numbers

Lower char

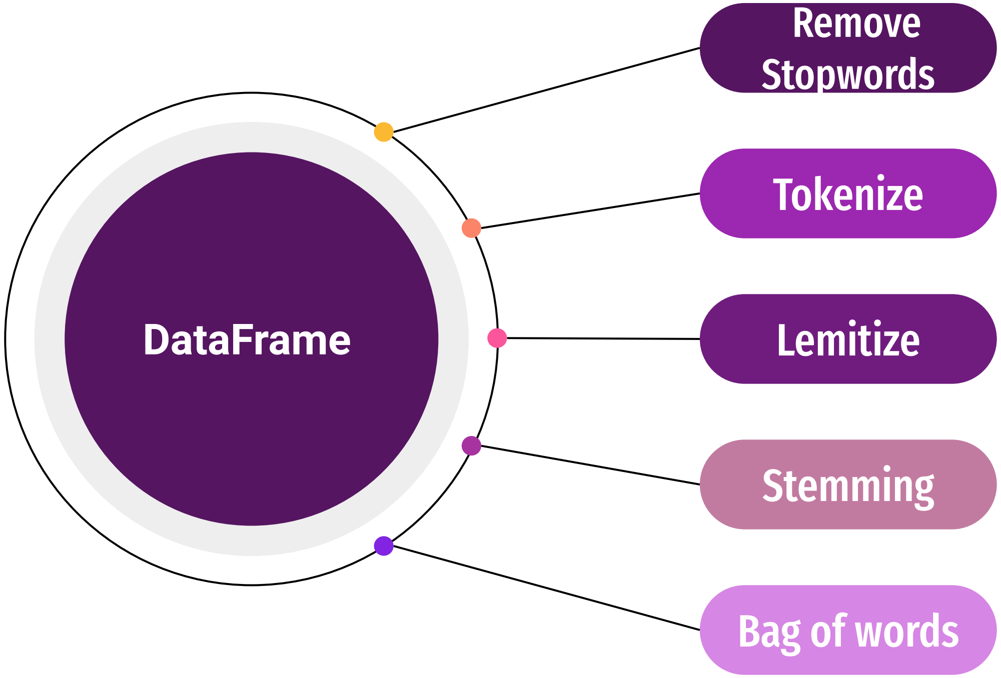
Transform to
lowercase characters

Data Preprocessing Example

‘PolySciMajor EPA chief doesn't think carbon dioxide is main cause of global warming and.. wait, what!? https://t.co/yeLvcEFXkC via @mashable’

polyscimajor epa chief doesnt think carbon dioxide is main cause of global warming and wait what urlweb via mashable

Data Normalization



DataFrame

Remove Stopwords

Tokenize

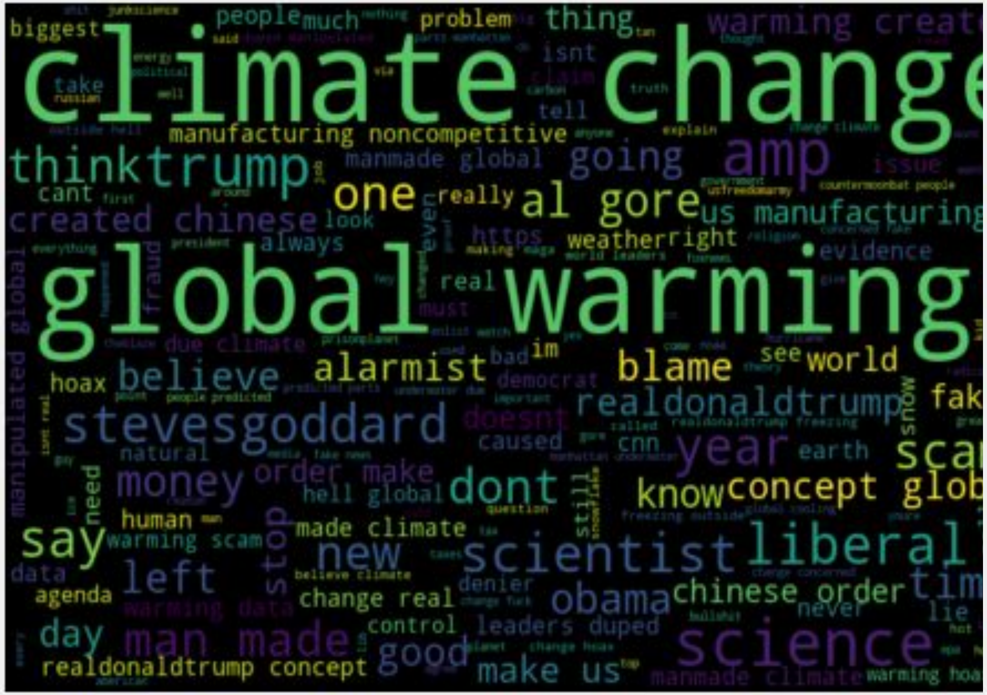
Lemitize

Stemming

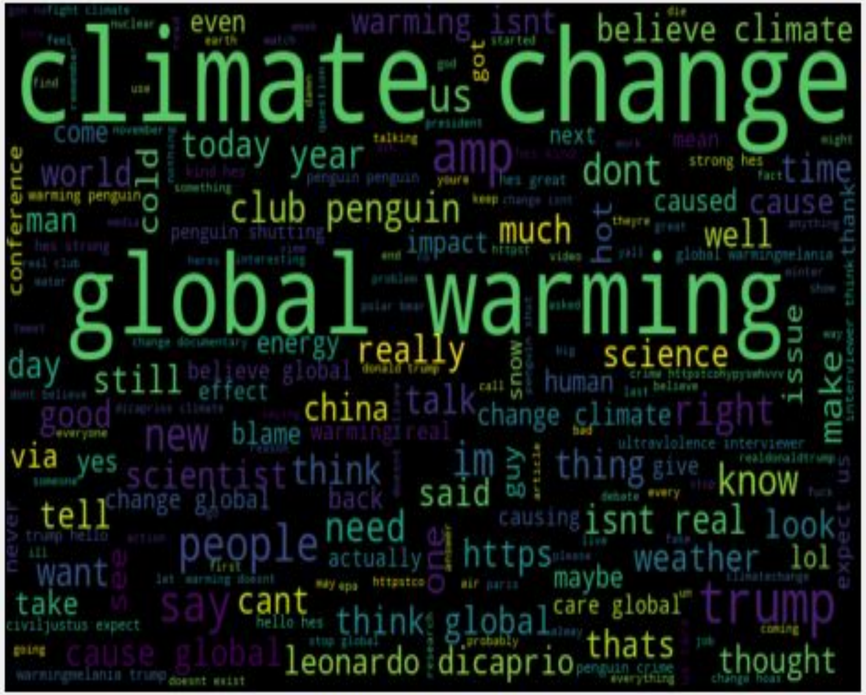
Bag of words

Further Text Exploration using wordcloud

Class -1

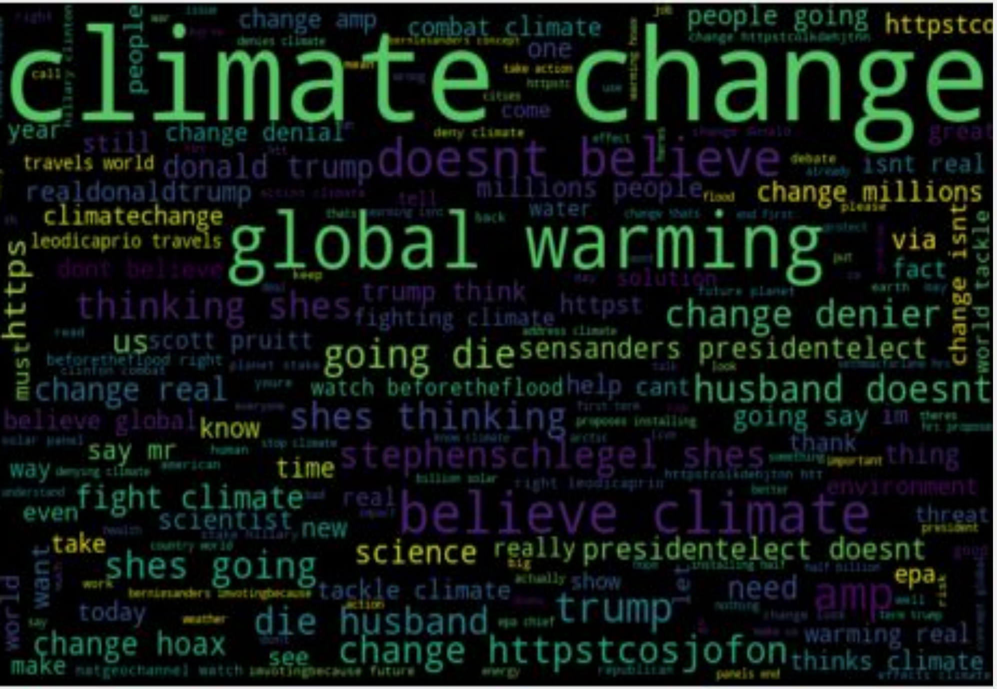


Class 0

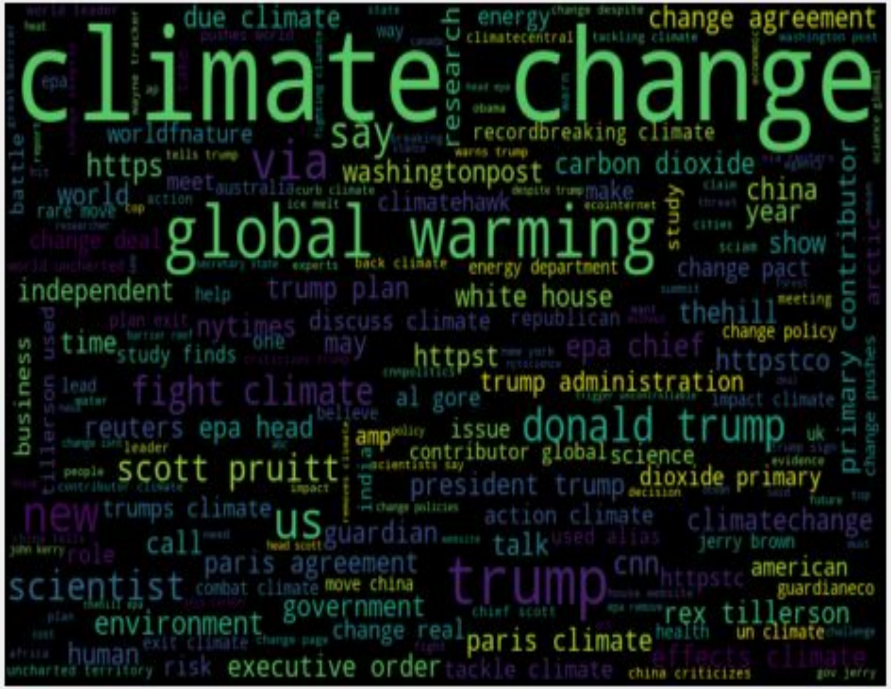


Further Text Exploration using wordcloud

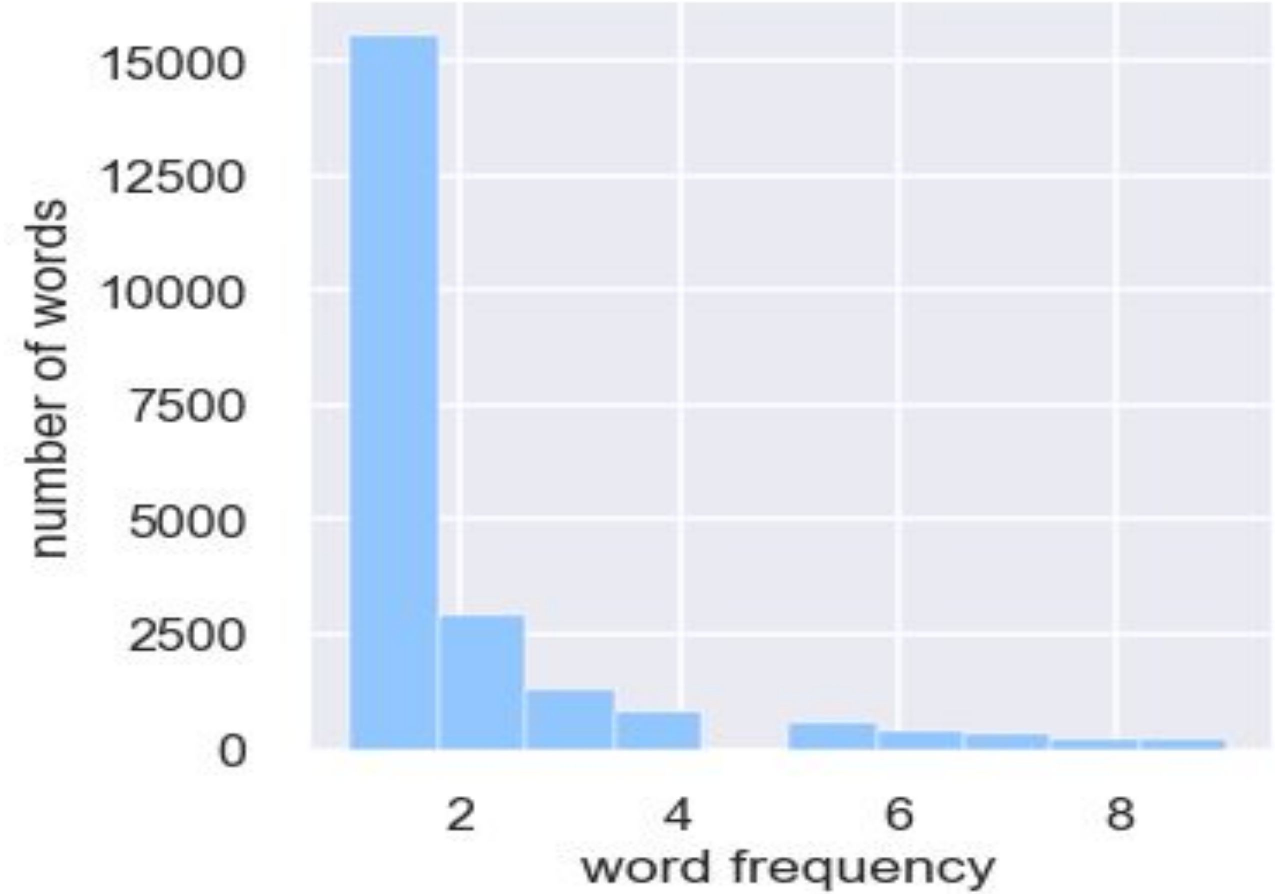
Class 1



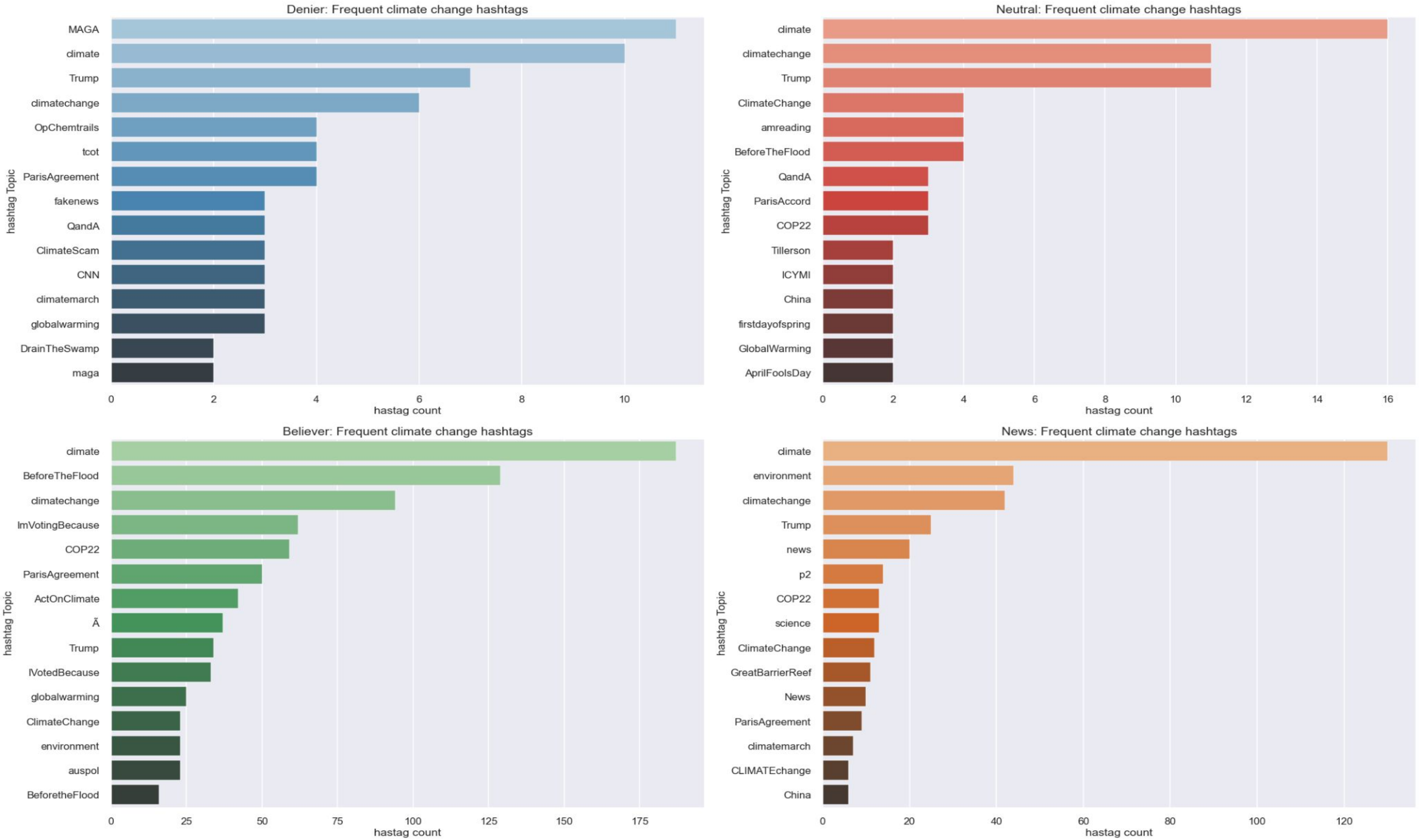
Class 2



Distribution of words Appearing < 10



Frequent HashTags for all the classes

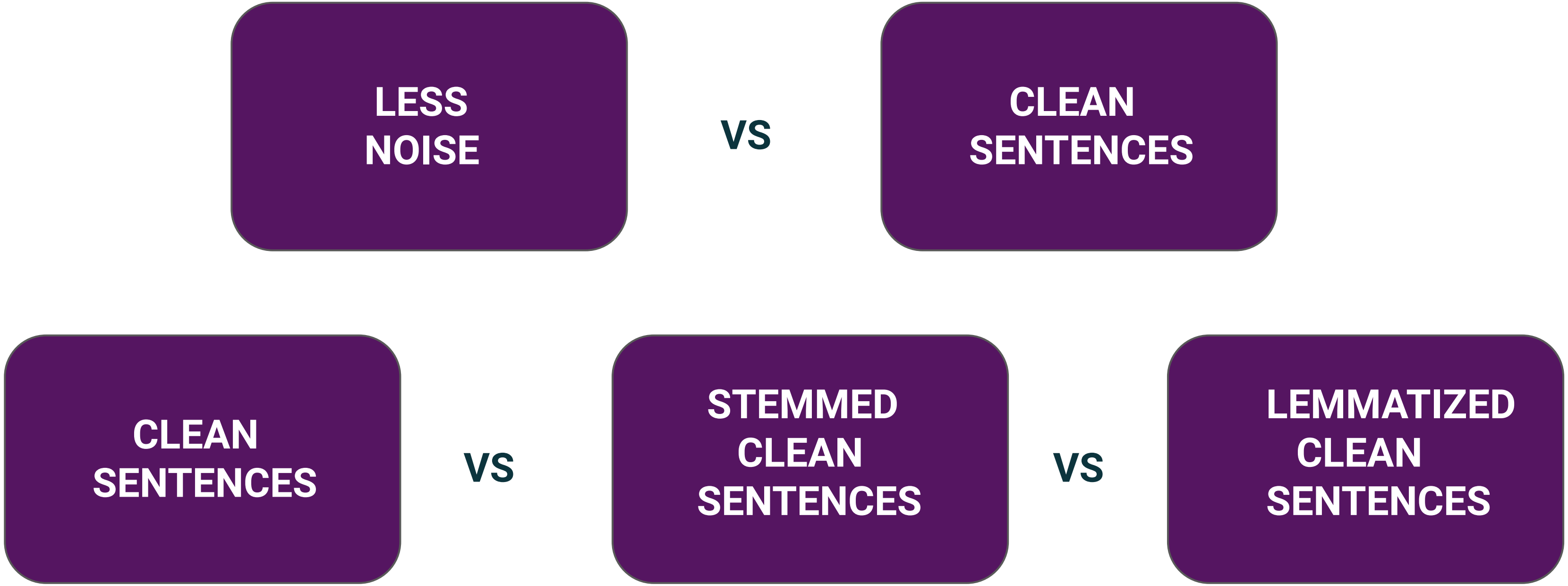

Denier: Frequent climate change hashtags
MAGA
climate
Trump
climatechange
OpChemtrails
tcot
ParisAgreement
fakenews
QandA
ClimateScam
CNN
climatemarch
globalwarming
DrainTheSwamp
maga
hashtag Topic
0
2
4
6
8
10
hastag count
Neutral: Frequent climate change hashtags
climate
climatechange
Trump
ClimateChange
amreading
BeforeTheFlood
QandA
ParisAccord
COP22
Tillerson
ICYMI
China
firstdayofspring
GlobalWarming
AprilFoolsDay
hashtag Topic
0
2
4
6
8
10
12
14
16
hastag count
Believer: Frequent climate change hashtags
climate
BeforeTheFlood
climatechange
ImVotingBecause
COP22
ParisAgreement
ActOnClimate
Ã
Trump
IVotedBecause
globalwarming
ClimateChange
environment
auspol
BeforetheFlood
hashtag Topic
0
25
50
75
100
125
150
175
hastag count
News: Frequent climate change hashtags
climate
environment
climatechange
Trump
news
p2
COP22
science
ClimateChange
GreatBarrierReef
News
ParisAgreement
climatemarch
CLIMATEchange
China
hashtag Topic
0
20
40
60
80
100
120
hastag count

Data Preprocessing

192
Rear words

192
Short Words

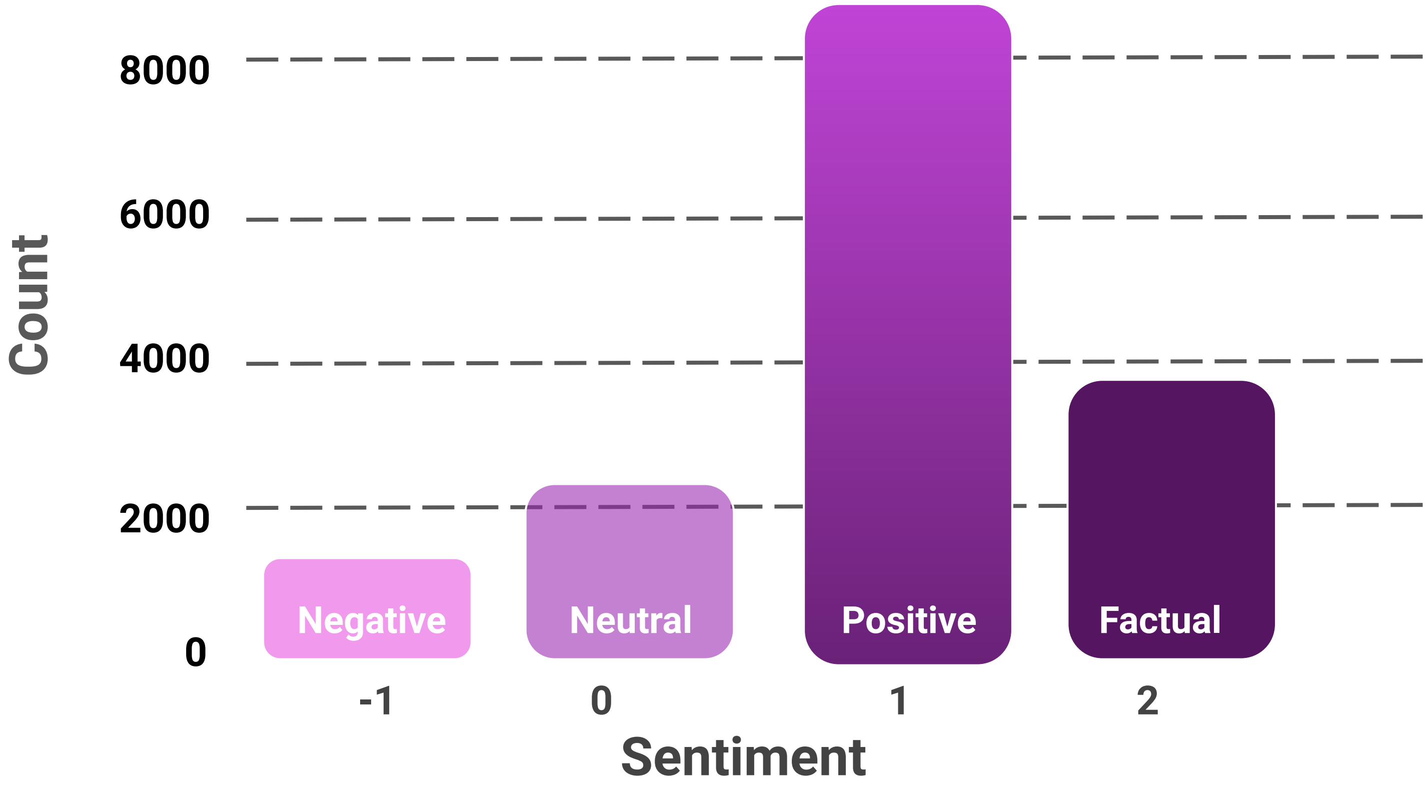
Experiments



LESS NOISE VS CLEAN SENTENCES

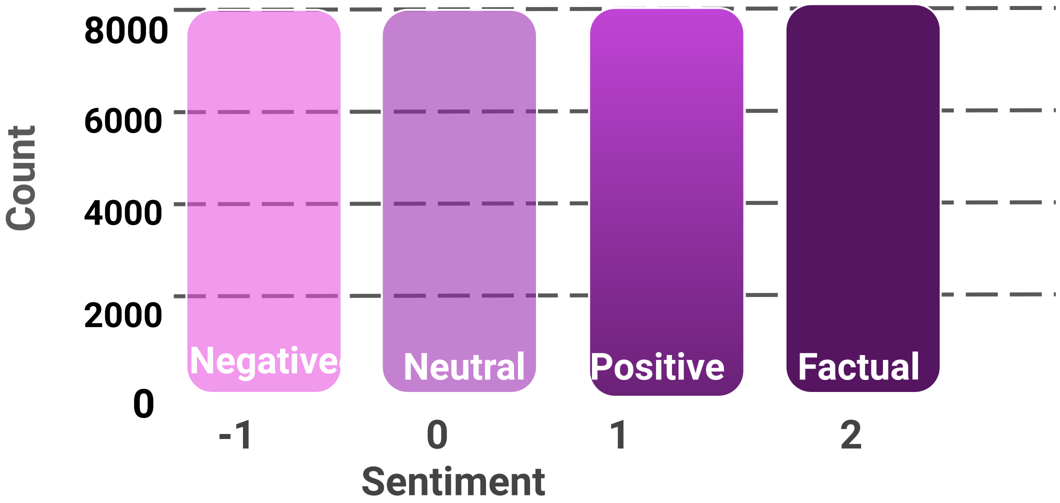
CLEAN SENTENCES VS STEMMED CLEAN SENTENCES VS LEMMATIZED CLEAN SENTENCES

Text Imbalance



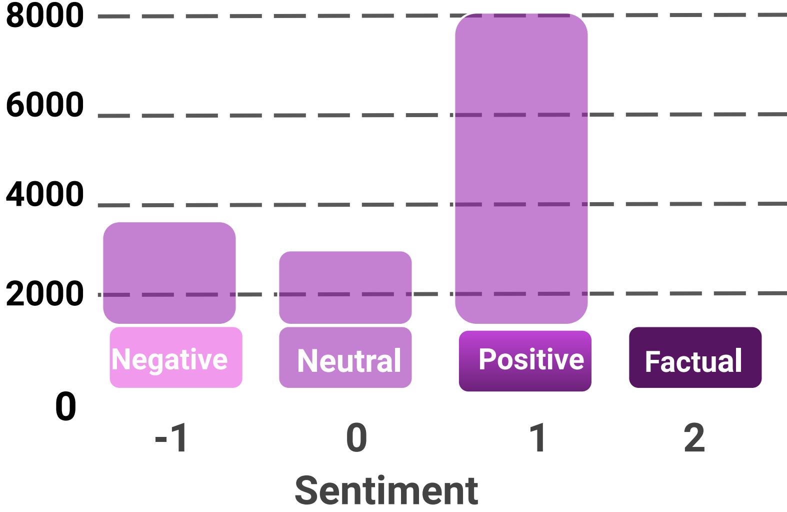
- Models become better at predicting one class over the others.
- Inherent predictive bias.

Upsampled Data



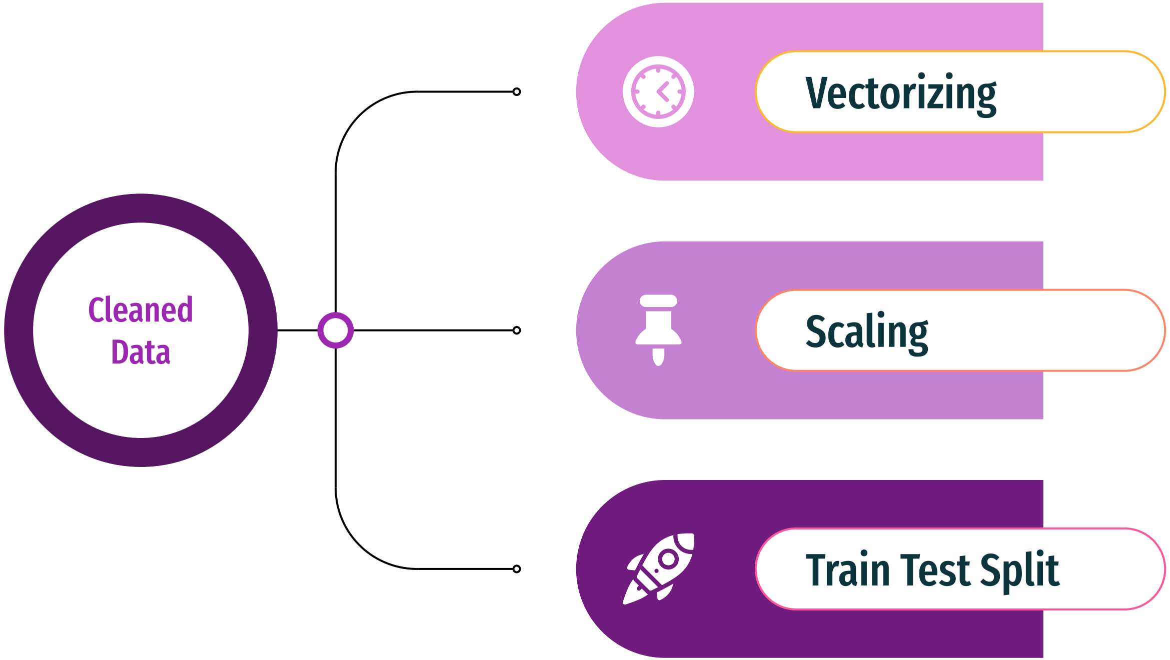
- Risk overfitting, can check against Test data.
- Generally better than downsampling

Downsampled Data



- Risk losing valuable information.
- Reduces dataset to a more manageable size.

Feature Text Engineering



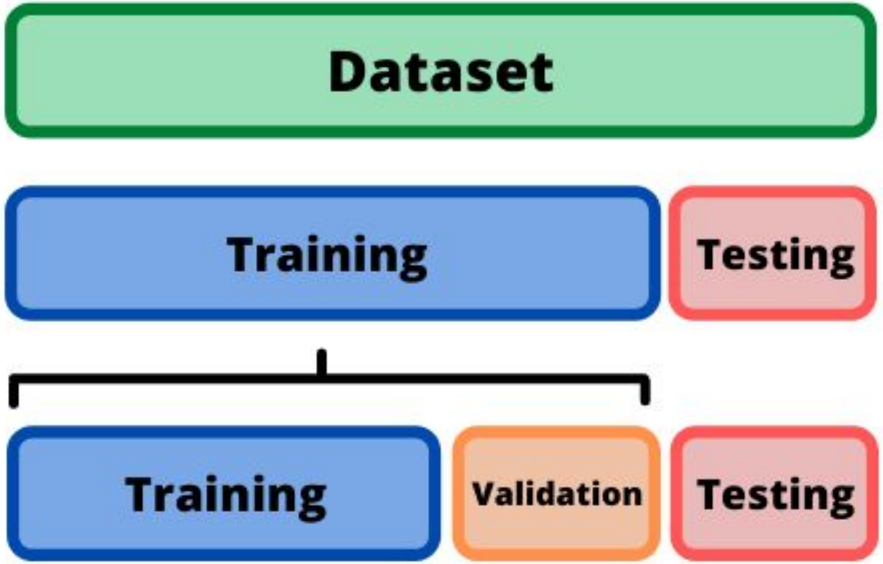
Cleaned Data

Vectorizing

Scaling

Train Test Split

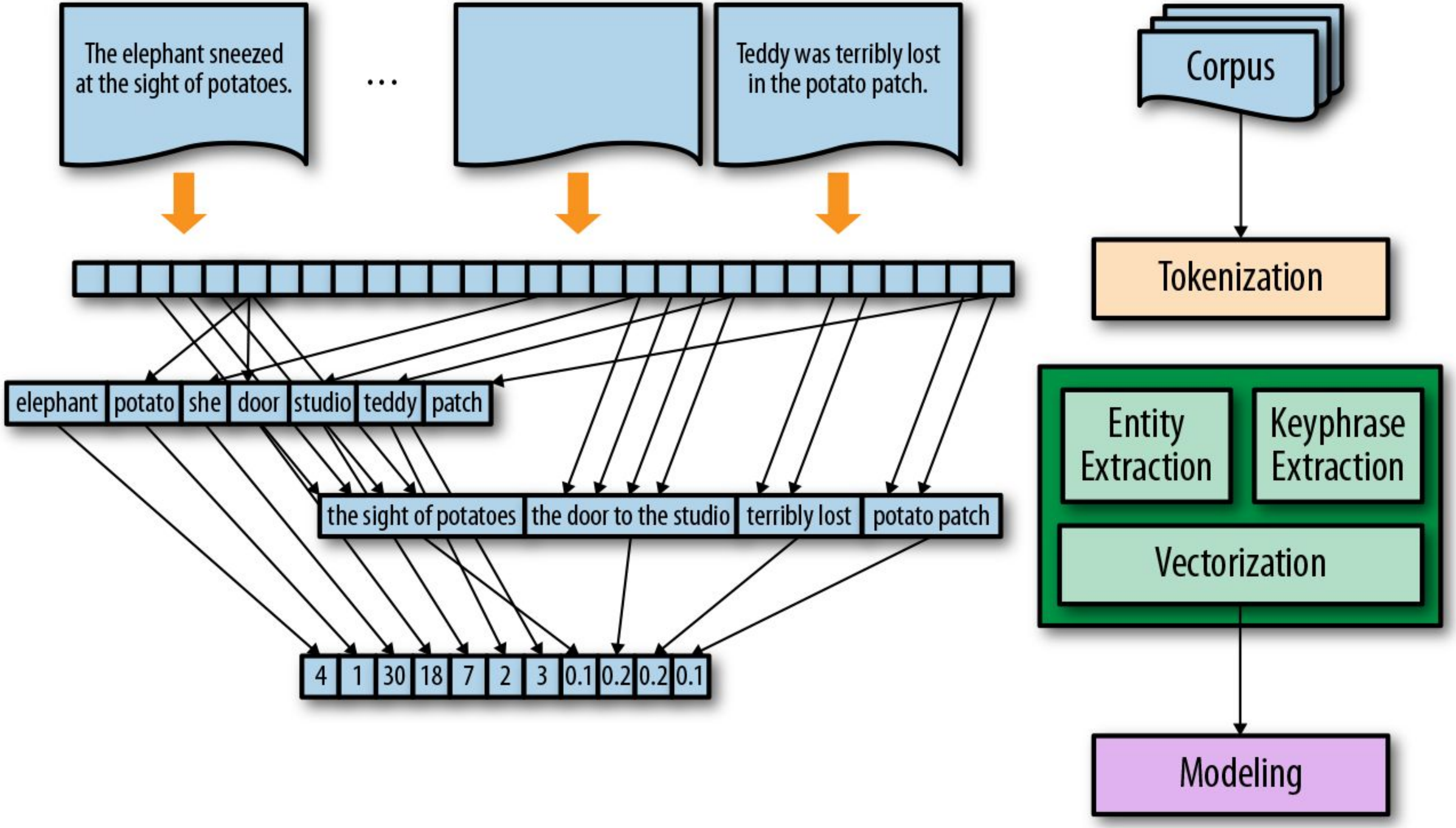
Train Test Split



Text Vectorisation

- A document term matrix is generated and each column represents an individual unique word.
- Each cell contains a weight value that signifies how important a word is for an individual text message or document.
- Different from the count vectorization in the sense that it takes into considerations not just the occurrence of a word in a single document but in the entire corpus.
- TF-IDF gives more weight to less frequently occurring events and less weight to expected events. So, it penalizes frequently occurring words that appear frequently in a document such as “the”, “is” but assigns greater weight to less frequent or rare words.

Text Vectorisation

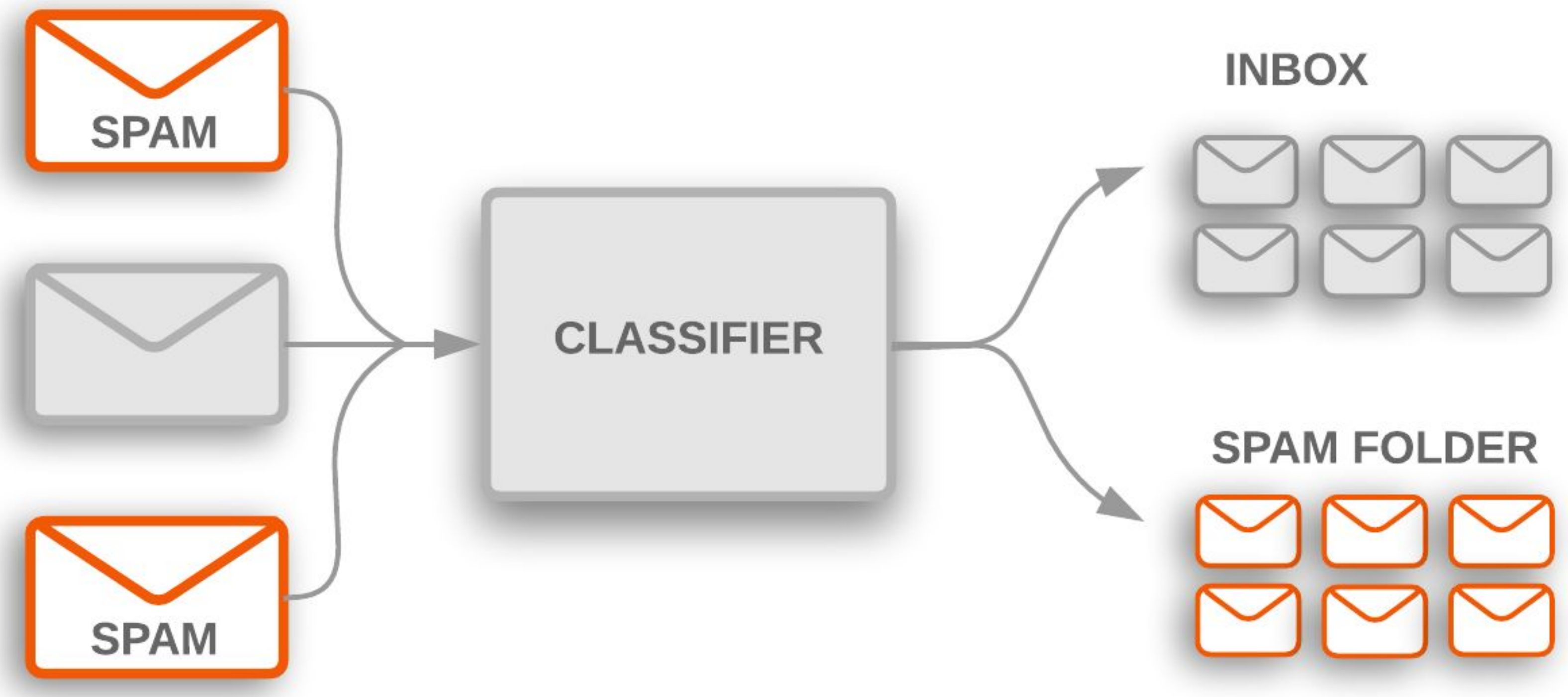


Data Scaling

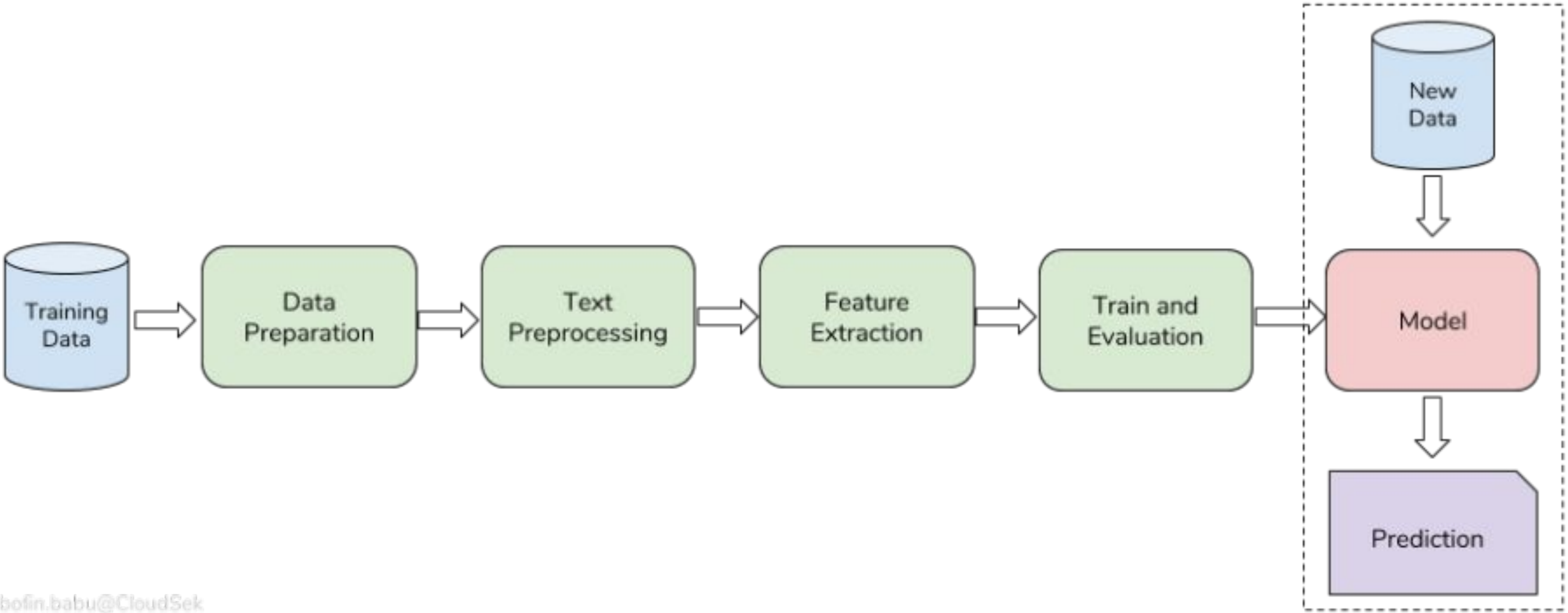
Min Absolute Scaler

- Scale each feature by its maximum absolute value.
- This estimator scales and translates each feature individually such that the maximal absolute value of each feature in the training set will be 1.0. It does not shift/center the data, and thus does not destroy any sparsity.
- This scaler can also be applied to sparse matrices.

Modelling

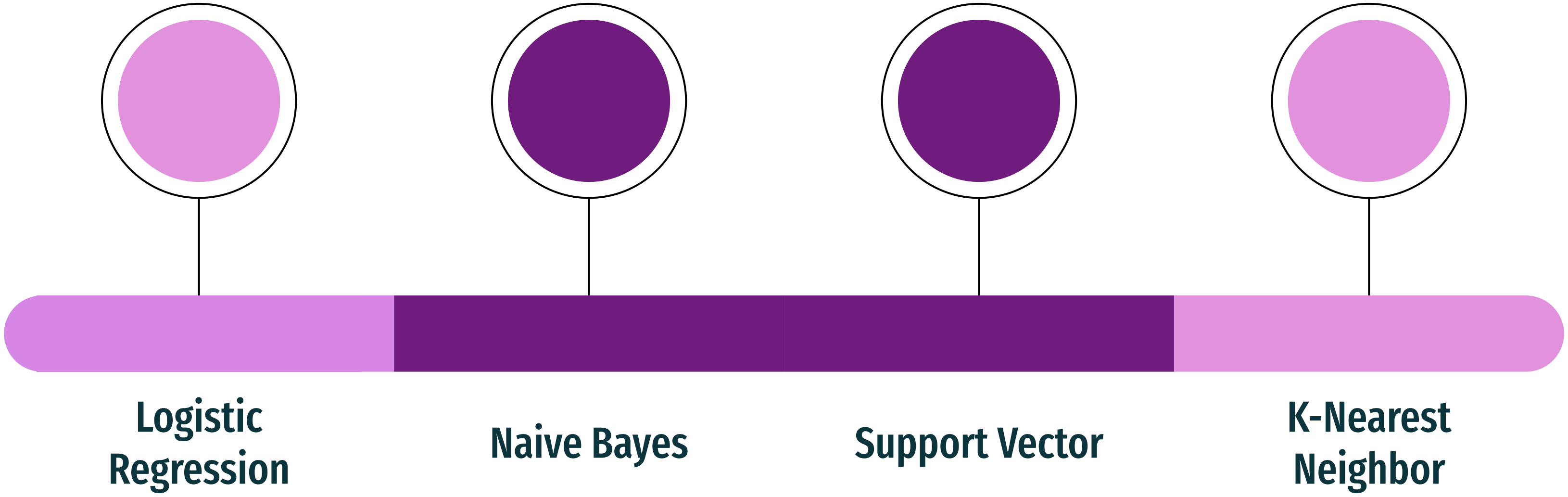


Modelling



Training Data
Data Preparation
Text Preprocessing
Feature Extraction
Train and Evaluation
New Data
Model
Prediction

Classification Algorithms



Logistic Regression

Naive Bayes

Support Vector

K-Nearest Neighbor