Discovering NLP and ML Fundamentals

Are you interested in learning how to teach machines to understand and generate human language? This presentation will introduce you to the concepts, history, and applications of Natural Language Processing and Machine Learning in the most stimulating way.

History of NLP

Language Technology and Revolution

NLP has rapidly developed since the 1950s, when experiments were conducted to simulate human-like responses to conversations.

The Emergence of Machine Learning

The introduction of machine learning models in the 80s led to the development of more accurate NLP systems.

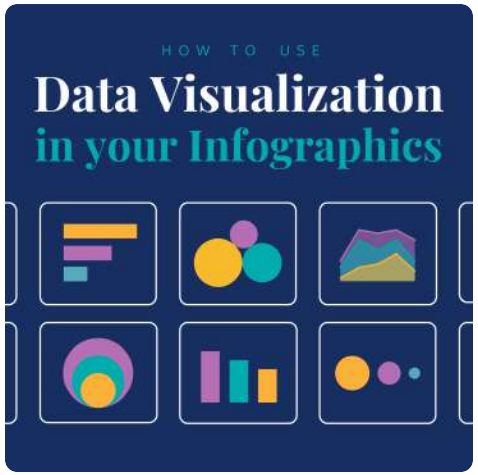
Resources and Digital Corpora

The internet and the creation of digital text collections have fueled the development of deep learning algorithms, significantly improving NLP.

Current Innovations

The latest advances in deep learning techniques and large language models have enabled solving previously unsolvable NLP problems, pushing the boundaries of what's possible.

Text Pre-Processing in NLP



Data Capture

NLP algorithms require clean and consistent data. OCR tools are used to digitize scanned text, while automatic speech recognition systems convert audio to text.

Data Cleaning

Text normalization is key, removing noise and irrelevant parts of the text, such as stopwords, and converting the corpus into a format suitable for language models.

Data Exploration

Exploratory data analysis helps researchers understand text patterns and distributions that are relevant to the models they are building.

Converting Text into Vectors

TF-IDF

A method that accounts for term frequency and inverse document frequency, reduces the dimensionality and compensates for rare terms. Popular in document retrieval and classification.

1

Bag of Words

A simple method, mapping each word in the corpus to a unique integer, which can lead to a high-dimensional and memory-intensive model.

2

3

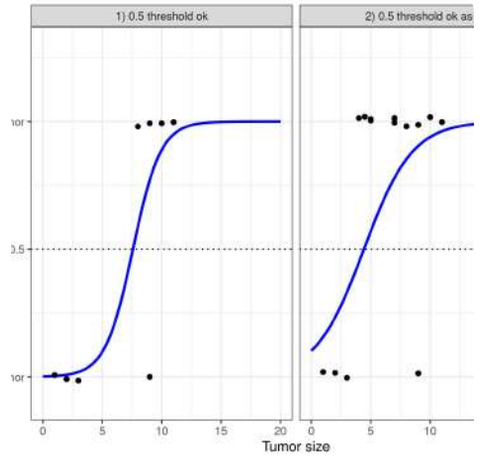
Word Embeddings

Using neural networks to learn distributed representations of words as dense vectors, leading to lower-dimensional and semantically-structured models.

Past Machine Learning Methods of NLP

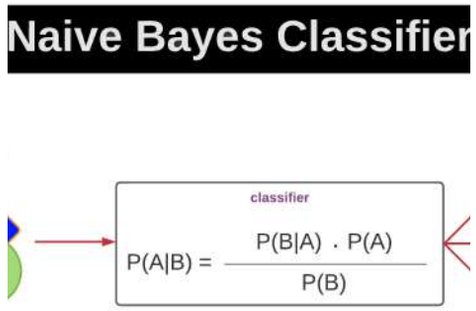
N-Gram Models

N-gram models are probabilistic models built by training on sequential text data. They count how often word sequences appear in the text and estimate the probability distribution of the word sequences.



Logistic Regression

Logistic regression is an algorithm that’s popular in text classification tasks. It uses a mathematical function that uses probability to generate a value between zero and one for any given numerical input. This value represents the probability of input belonging to a particular class.



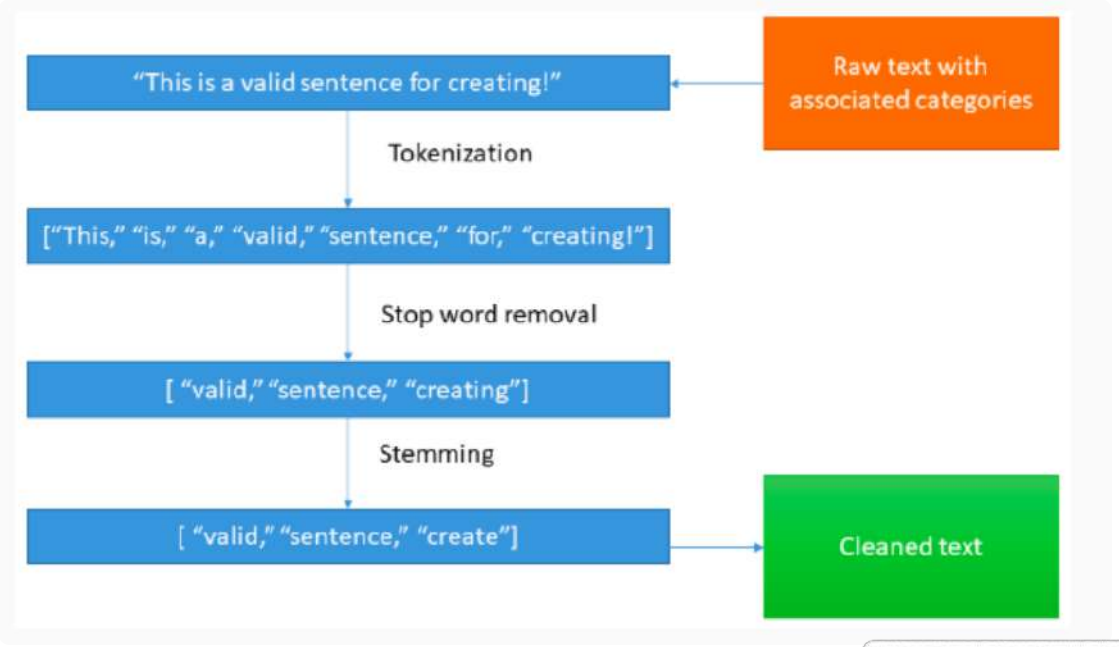
Naïve Bayes

Naïve Bayes is a supervised algorithm that uses probability to predict the likelihood of a class being the outcome. The Naïve algorithm uses Bayes’ theorem together with the (naïve) assumption that features in the data are independent of each other.

How to Build a Classifier

1 Data Preparation

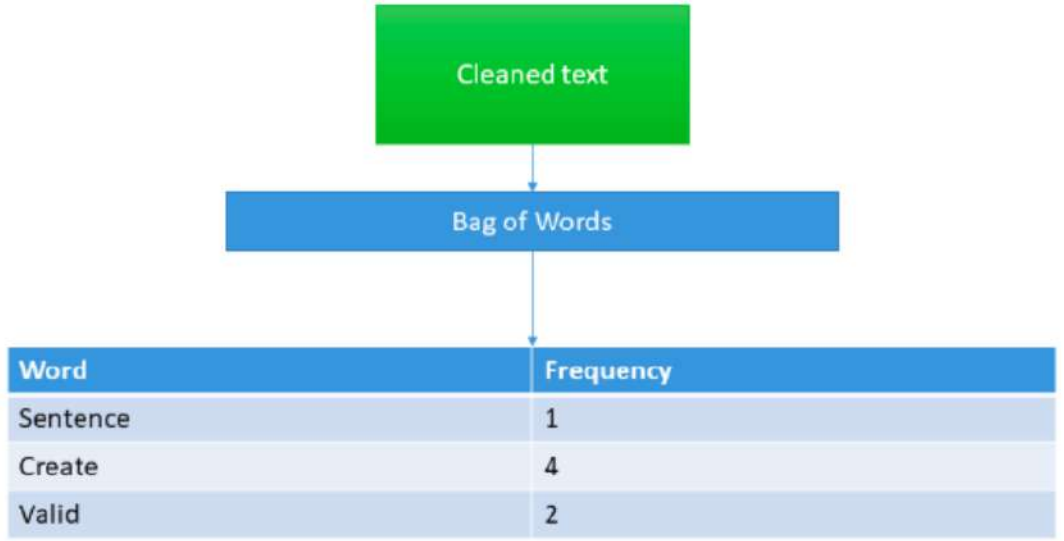
Choose and preprocess the data based on the classification task while preserving its proportionality.



How to Build a Classifier

1 Text Vectorization

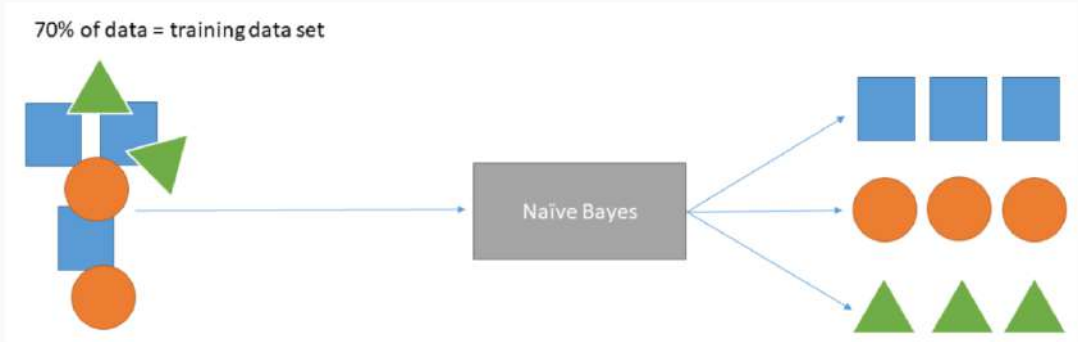
With text vectorization, you must convert the documents in your corpus into a numeric representation. This helps the ML algorithms to process the data as they work better with numbers than text.



How to Build a Classifier

1 Train the Model

To build the supervised classifier, you must identify the algorithm you want to use. This overview will use the naïve Bayes algorithm. For each document in your corpus, you now have its associated vector representation and the target variable.



How to Evaluate a Classifier

Accuracy

- The ratio of correctly classified instances to the total number of instances.

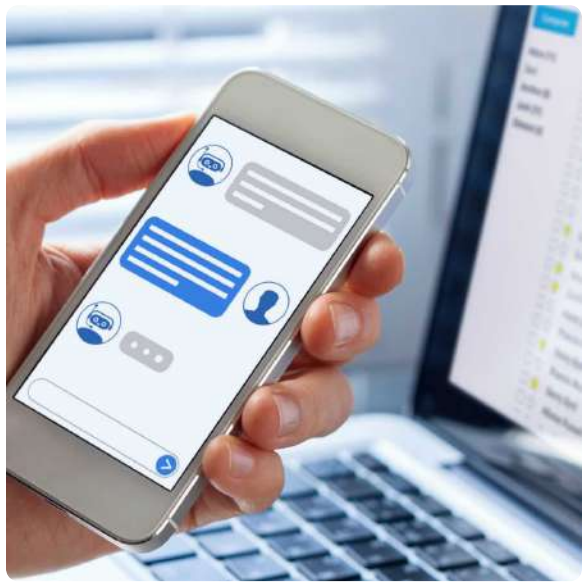
Precision and Recall

- Precision is the number of true positive predictions divided by the total number of positive predictions.
- Recall is the number of true positive predictions divided by the number of positive instances in the test set.

F1 Score and AUC

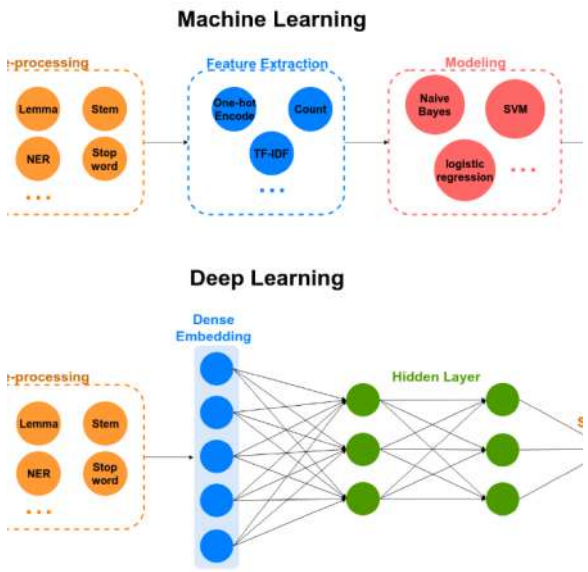
- The harmonic mean between Precision and Recall, providing a balance between both metrics.
- The area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate against the False Positive Rate of the model.

Applications of NLP



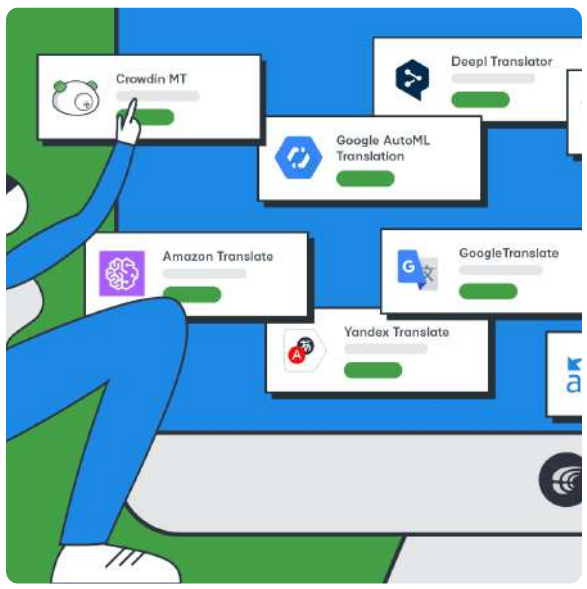
Virtual Assistants

NLP is at the core of digital assistants such as Siri or Alexa, enabling human-like voice interactions and intelligent responses.



Sentiment Analysis

Classifying the polarity of text based on emotional or critical expressions is a common task, with applications in customer feedback, social media, and business intelligence.



Machine Translation

By learning the patterns and structures of different languages, machine translation models are able to translate entire texts from one language to another.

Chatbots

The automated and dynamic nature of rule-based and machine learning-based chatbots makes them useful for customer service, health, education, and entertainment.

Recommended Reading

Ultimate Guide to Understand and Implement Natural Language Processing (with codes in Python)

Link: https://www.analyticsvidhya.com/blog/2017/01/ultimate-guide-to-understand-implement-natural-language-processing-codes-in-python/

Next: Large Language Models

1

What are Large Language Models?

NLP models with billions of parameters learned from vast amounts of diverse text data using deep learning architectures.

2

Why are they Important?

Large language models are the driving force behind natural language processing breakthroughs seen in recent years, such as BERT, GPT-3, or T5. They have enabled improvements in speech recognition, language translation, and reading comprehension tasks.

3

What are the Challenges?

Training large language models is a computationally intensive task, requiring massive processing power and data centres. These models are opaque and difficult to interpret, leading to debates on ethical, legal, and social issues.