

Artificial Intelligence

ACCA7026

Natural Language Processing – Introduction

Dr. Seena Joseph

Natural Language Processing (NLP)

Debater

Understand the situation, build a position, Organize the content and respond to the opponent

Example: **IBM Project Debater**

How this is possible ?

NLP : A field of AI, that focus on enabling computers to understand, interpret and generate human language.

Human language is unstructured and have several meaning.

Computers deals with this messiness by considering one sentence at a time (sentence segmentation)

Then divide the information into small chunks (Tokens) that can be individually classified.

Natural Language Processing (NLP)

Key elements are entity, Relationship and concepts.

Tokenized texts are categorised into Entity and Relationship.

Entity : An entity is a noun representing a person, place, or thing. It's not an adjective, verb, or other article of speech.

Example: What is the price of iPhone 16 ?

Entity : product (iPhone 16)

Show me hotels in Paris from December 10 to December 15 with a pool.

Entity : Location (Paris); Check in Date (December 10) Checkout Date (December 15) and Amenity (Pool)

Natural Language Processing (NLP)

Relationship : It is a group of two or more entities that have a strong connection to one another.

The price of iPhone 16 is £900.

Relationship – Price (£900) and product (iPhone 16) .

Smith's Birthday is on Jan 2nd

Relationship – Person (Smith) and Date (Jan 2nd)

Concept : it is something implied in a sentence but not actually stated. This is trickier because it involves matching ideas rather than the specific words present in the sentence.

It is an abstract idea, that represents a general theme or group.

Concepts helps NLP to understand the bigger picture or context of the situation/dialogue.

Example : **Concept** : Mode of transport.

- *I will take bus to work.*
- *Book me a flight to London.*

Natural Language Processing (NLP)

Emotion Detection and Sentiment Analysis

Emotion : Happiness, anger, fear

AI can be trained to detect and classify emotions.

Sentiment analysis – not an emotion, but a measure of the strength of an emotion.

Sentiment – a sliding scale between positive and negative and neutral in between.



Natural Language Processing (NLP)

Human language is full terms with different meanings

I went to the dock to ship my parcel.

I went to the station to ship my parcel.

How an AI system recognise these context and interpret how the parcel will be transported ?

Natural Language Processing (NLP)

Chatbots

- Programs that employ NLP to simulate conversations with humans.
- Programmed to ask a limited set of questions.
- It works with small data.

Example : A chatbot for assisting resetting password.

How can I reset my password ? This is a straightforward way to ask questions.

However, this can be answered in several ways..

1....?

2...?

This is achieved by Classifiers – it can map different ways of asking questions to a very small set of answers.

Examples for Chatbots services ?

Natural Language Processing (NLP)

Chatbot

Frontend – messaging part or user interface or

Backend – classifier

- intents, entities and dialog

A chatbot understand a question by segmenting a sentence into parts and relating those parts into things in memory.

The role of a chatbot is to identify **entities** and **intents** then uses these to trigger a **dialogue**.

Intent – The purpose . Why is a user contacting a chatbot ?

Example : to get direction ?

: to reset passwords ?

: to track a parcel ?

Dialogue – a flowchart IF THEN tree structure...

Activity

If you are asked to design a chatbot for a restaurant chain.

What would be the possible intents?

Natural Language Processing (NLP)

Dialogue – Represents each possible word or phrase

- The matched response for the chatbot.
- A many subsequent replies

It is a very complex flowchart representation

So chatbot software condenses each moment of the conversation into a **node**. A node contains a statement by the chatbot and a long, expandable list of possible replies.

Many possible questions would be mapped onto a small number of possible answers until the conversation is ended.

Natural Language Processing (NLP)

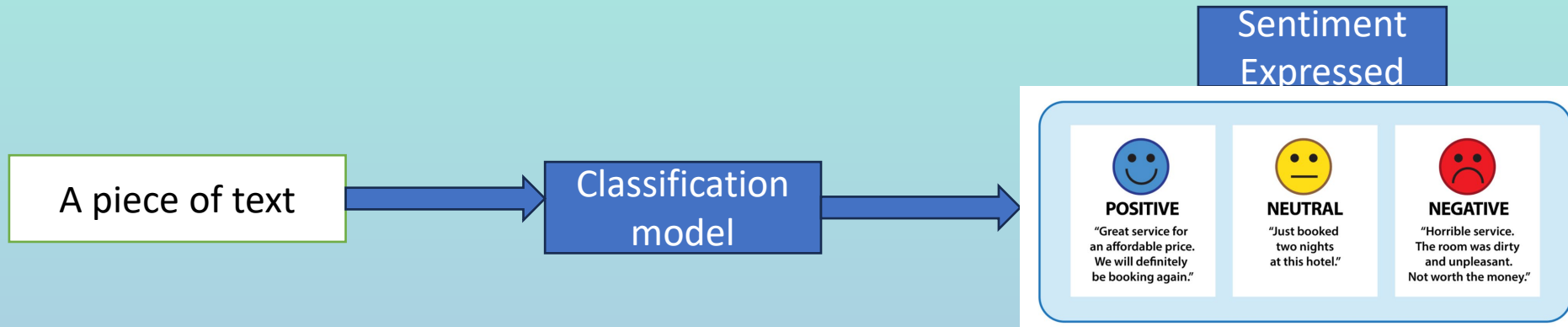
E-commerce customer support chatbot

User message	Intent	Entity	Dialogue
"Where is my order?"	Order tracking	None	Please provide your order number.
Can I return the Christmas tree that I ordered last week?	Initiate return	Christmas tree, Date, Last week	Yes of course, could you share the product number. I will guide you through the process.
?			
?			
?			

Natural Language Processing (NLP)

Why NLP ?

1. **Sentiment analysis:** it is the process of classifying the emotional intent of text.



Sentiment analysis is used to classify customer reviews on various online platforms as well as for niche applications like identifying signs of mental illness in online comments.

Natural Language Processing (NLP)

Why NLP ?

2. **Language Translation** : The process of transforming spoken or written language from one language to other. Machine translation (MT) tools automate translation between languages. These models are used to improve communication between people.

Example: Automate the process of providing answers to customer queries by building virtual assistance to translate inbound questions and responses from customers into other languages in real time.

Social media platforms such as Facebook automatically translates posts and comments using MT to remove language barriers and enable global communication.

Voice based online searches using Siri, Alexa, Google Assistant.

Natural Language Processing (NLP)

Why NLP ?

3. **Named entity recognition (NER)** : aims to locate and extract entities in a piece of text into predefined categories such as personal names, organizations, locations, and quantities. The input to such a model is generally text, and the output is the various named entities along with their start and end positions.

Example : in an automated customer query service, these models can be used to extract the name of a product from a customer's inquiry and then utilising that name to tell the customer about the product's price, qualities, and availability.

Also be used to identify the intent of a customer, for example 'cancel' or 'refund' and direct the query to the relevant department.

Natural Language Processing (NLP)

Why NLP ?

3. Grammatical error correction : Encode grammatical rules to correct the grammar within text. The model is trained on a ungrammatical sentence as input and a correct sentence as output.

Example : Grammarly, MS Word.

4. Text Generation : Aims to produce text that is similar to human language. Such models can be fine-tuned to produce text in different genres and formats including tweets, blogs, and even computer code.

5. Toxicity Classification – it is a branch of sentiment analysis where the aim is not just to classify hostile intent but also to classify particular categories such as threats, insults, obscenities, and hatred towards certain identities.

Natural Language Processing (NLP)

Why NLP ?

6. Information retrieval : Finds documents that are relevant to the query. The goal is not to answer a particular query but to retrieve, from a collection of documents that may be numbered in the millions, a set that is most relevant to the query.

7. Summarization : is the task of shortening text to highlight the most relevant information. This process Condense a piece of text into a shorter version while retaining the key information and main ideas.

Extractive summarisation: Selects important sentences or phrases directly from the original text without modifying them.

Example:

If the text is: "*Natural Language Processing enables machines to understand human language. It has many applications like sentiment analysis, translation, and summarization.*"

The extractive summary could be: "*Natural Language Processing enables machines to understand human language. It has applications like summarization.*"

Natural Language Processing (NLP)

Why NLP ?

Abstractive Summarization: Generates new sentences that capture the essence of the original text. It uses advanced models like transformers to rewrite the content in a concise form.

Example: From the same text above, an abstractive summary could be: "*NLP helps machines comprehend human language, supporting tasks like sentiment analysis and summarization.*"

Natural Language Processing (NLP)

How Does Natural Language Processing (NLP) Work?

NLP models aims to find relationship between constituent parts of the language- for example the letters, words, and sentence found in a text dataset.

Text preprocessing – it is important to convert the textual data in a format that the model can understand. The common pre-processing techniques are : **Stemming and lemmatization, sentence segmentation, stop word removal, and Tokenization.**

Stemming and lemmatization : Stemming is the process of converting the word into the basic/root form known as stem. The purpose of stemming is to group similar words with the same root meaning under a common representation.

Natural Language Processing (NLP)

Stemming and Lemmatization:

Lemmatization is a more formal way to find roots by analyzing a word's morphology using vocabulary from a dictionary. It reduces the words to their lemma or dictionary root form.

Word	Stem	Lemma (Base Form)
Running	Run	Run
Runs	Run	Run
Happily	Happi	Happy
Happiness	Happi	Happiness
Studies	Studi	Study
Studying	Studi	Study

Natural Language Processing (NLP)

Sentence segmentation: It breaks a large piece of text into linguistically meaningful sentence units.

The goal is to identify the sentence boundaries, usually marked by punctuations like

- ? Question marks

- . Periods

- ! Exclamation Marks

Example :

Input Text:

"Dr. Smith arrived at 10 a.m. The meeting was productive! What did he say about the project?"

Segmented Sentences:

Dr. Smith arrived at 10 a.m.

The meeting was productive!

What did he say about the project?

Natural Language Processing (NLP)

Stop word removal : It aims to remove most commonly occurring words that don't add much information to the text or contribute to the analysis of the content. For example, “the”, “a”, “is”, “an”, and “of” .

Common stop words:

Articles: "the," "a," "an“

Conjunction : "and," "or," "but“

Prepositions: "in," "on," "at"

Pronouns: "he," "she," "they

Example:

Input Text: "*The cat sat on the mat and looked at the sky.*“

After Stop Word Removal:

"cat sat mat looked sky“

Here, stop words like "the," "on," "and," "at" have been removed, leaving only the meaningful words.

Natural Language Processing (NLP)

Tokenization

Tokenization splits text into individual words and word fragments.

It is to split the larger body of text (sentence, paragraph or documents) into smaller tokens.

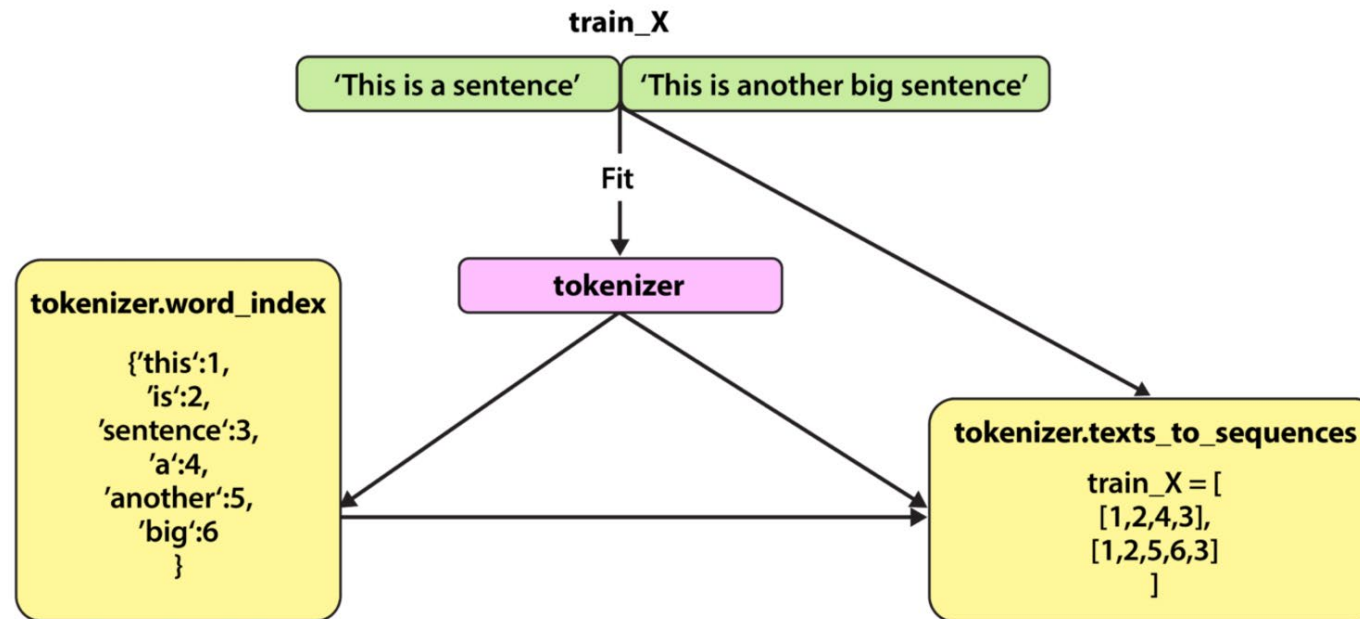
These tokens can be word, phrases or character depending on the level of granularity.

The result generally consists of a word index and tokenized text in which words may be represented as numerical tokens for use in various deep learning methods.

Natural Language Processing (NLP)

Tokenization

TOKENIZERS



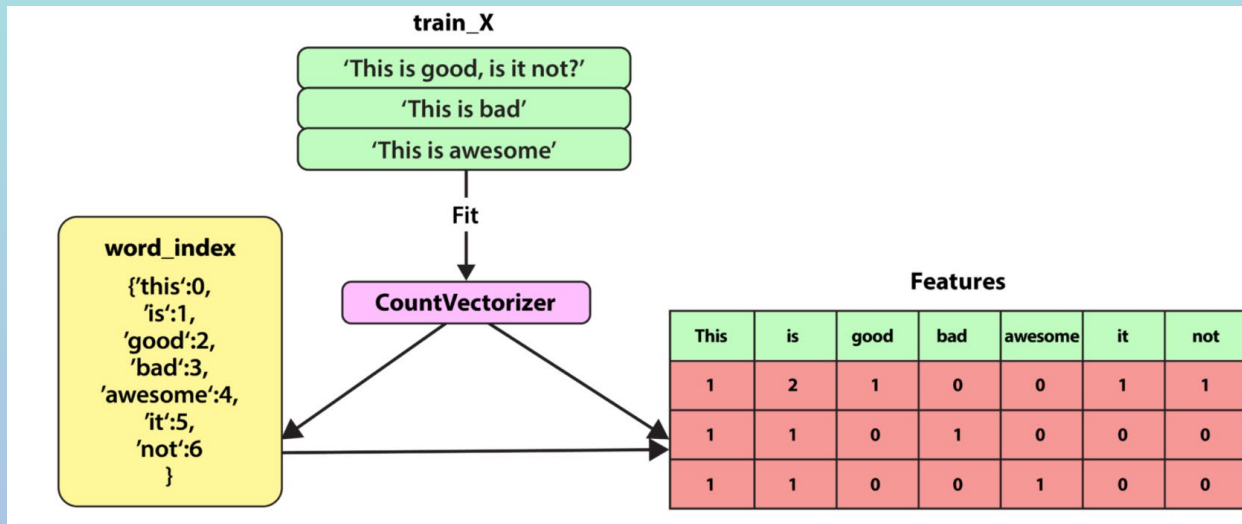
Natural Language Processing (NLP)

Feature Extraction

In NLP the feature extraction transform the text into a numerical representations that ML/DL algorithms can process.

Common Feature extraction techniques : Bag of Words and TF-IDF (Term Frequency Inverse Document Frequency)

Bag of Words: It counts the number of times each word or n-gram (combination of n words) appears in a document.



Natural Language Processing (NLP)

Feature Extraction

TF-IDF (Term Frequency Inverse Document Frequency):

TF-IDF, weights each word by its importance. Significance of the word is based on:

- **TF(word in a document)** – How important the word in the document?

TF (word in a document) = Number of occurrences of that word in document / Number of words in document.

- **IDF** – this measures how important is the word in the whole corpus? Or how unique the word across all.

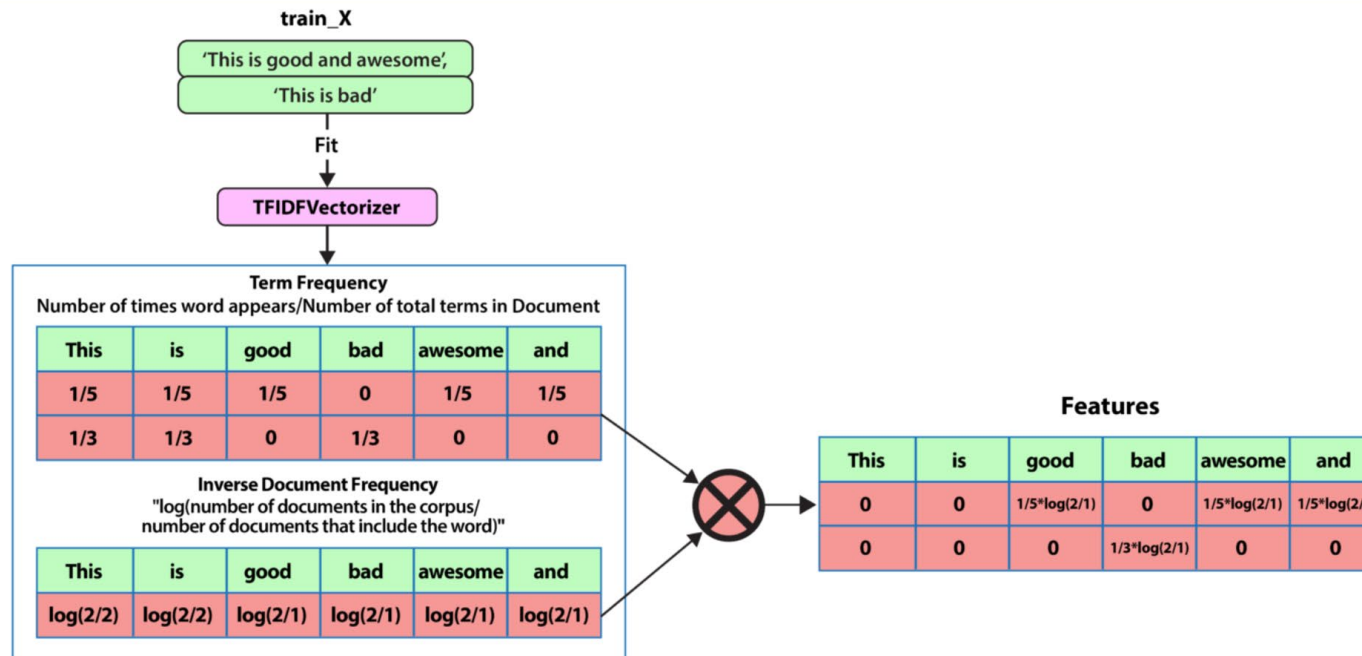
IDF(word in a corpus)= $\log(\text{number of documents in the corpus} / \text{number of documents that include the word})$

Natural Language Processing (NLP)

Feature Extraction

- TF-IDF score will be high if the word is rare and low if the word is common across the corpus. The TF-IDF score of a term is the product of TF and IDF.

TOKENIZERS: TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY (TF-IDF)



Natural Language Processing (NLP)

Word Embedding – numerical representation of words that capture the meaning.

Word2Vec and GLoVe are two common techniques for word embedding.

Natural Language Processing (NLP)

Modelling: After data is preprocessed, it is fed into an NLP architecture that models the data to accomplish a variety of tasks such as text classification, machine translation, sentiment analysis or language generation.

In this stage the numerical features extracted can be fed into different ML/DL models to analyse, understand and generate human language.

For example : For example, for classification, the output from the TF-IDF vectorizer could be provided to **logistic regression** or **decision trees**.

Deep neural networks typically work without using extracted features, although we can still use TF-IDF or Bag-of-Words features as an input.

Or **Language models** can be used,

Natural Language Processing (NLP)

Natural Language Processing Models

- Eliza
- Tay
- BERT
- Generative Pre-Trained Transformer (GPT)
- Language Model for Dialogue Applications (LaMDA)
- Mixture of Experts (MoE)

Natural Language Processing (NLP)

NLP Practical Task

Task Objective : Classify customer reviews as either "positive" or "negative" based on the sentiment expressed in the text using deep learning techniques.

Steps:

- **Choose a dataset** Choose a dataset that contains text reviews with sentiment labels (positive or negative). You can find datasets from sources like Kaggle, Twitter, or Yelp. Ensure that the dataset has enough reviews to train a deep learning model.
- **Text preprocessing**
- **Feature Representation**
- **Model building** (using any models such as CNN, RNN, LSTM, GRU or ensemble)
- **Model Training**
- **Evaluation**

Natural Language Processing (NLP)

Case Study

Article:

"NLP techniques for automating responses to customer queries: a systematic review" by P. A. Olujimi and A. Ade-Ibijola.

Extra Resources

[Natural Language Processing \(NLP\) \[A Complete Guide\]](#)

[How to make Chatbot with Google Dialogflow: Key Insights](#)

Reference

[1] P. A. Olujimi and A. Ade-Ibijola, "NLP techniques for automating responses to customer queries: a systematic review," *Discover Artificial Intelligence*, vol. 3, no. 1, 2023. DOI: 10.1007/s44163-023-00065-5.