Tokenizer function

can u explain what this function does:

Your function:

```
def tokenize(batch):
    return tokenizer(
        batch["tweets"],
        truncation=True,
        padding="max_length",
        max_length=128
    )

dataset = dataset.map(tokenize, batched=True)
```

Your function:

```
def tokenize(batch):
    return tokenizer(
        batch["tweets"],
        truncation=True,
        padding="max_length",
        max_length=128
    )

dataset = dataset.map(tokenize, batched=True)
```

1️⃣ What `.map()` does

`dataset.map()` in Hugging Face applies a function to **every example in your dataset**.

- `batched=True` → the function gets a **batch of examples** at once (like a mini-list), not a single example.
- `batch` is a dictionary where keys are column names, values are **lists of values**.

Example:

Suppose your dataset has 3 tweets:

tweets	label
"I love Borderlands!"	2
"This game is boring."	0
"Neutral review of the game."	1

With `batched=True`, the `batch` looks like:

```
batch = {
    "tweets": [
        "I love Borderlands!",
        "This game is boring.",
        "Neutral review of the game."
    ],
    "label": [2, 0, 1]
}
```

2️⃣ What the `tokenizer` does

`tokenizer(batch["tweets"], ...)` converts **raw text into numbers** (input IDs) that the model understands.

Here's what each argument does:

Argument	Purpose
`batch["tweets"]`	The raw text (list of tweets)
`truncation=True`	Cut off tweets longer than `max_length`
`padding="max_length"`	Pad shorter tweets to `max_length` with zeros
`max_length=128`	All sequences become exactly 128 tokens

The tokenizer converts each string into a dictionary:

```
{
  'input_ids': [...],        # numerical IDs for each token
  'attention_mask': [...]    # 1 for real tokens, 0 for padding
}
```

---

3 What your `tokenize` function returns

```
{
  "input_ids": [
      [101, 1045, 2293, 2342, 999, 102, 0, 0, ...],  # "I love
Borderlands!"
      [101, 2023, 2190, 2003, 6659, 1012, 102, 0, ...],  # "This
game is boring."
      [101, 8089, 4605, 1997, 1996, 2342, 1012, 102, ...] # "Neutral
review..."
  ],
  "attention_mask": [
      [1,1,1,1,1,1,0,0,...],  # same length as input_ids
      [1,1,1,1,1,1,1,0,...],
      [1,1,1,1,1,1,1,1,...]
  ]
}
```

- **`input_ids`** → what the model sees
- **`attention_mask`** → tells the model which tokens are real vs padding

---

4 How `.map()` updates the dataset

After mapping, your dataset now has **new columns**:

input_ids	attention_mask	label
[101, 1045, 2293, ...]	[1, 1, 1, ...]	2
[101, 2023, 2190, ...]	[1, 1, 1, ...]	0
[101, 8089, 4605, ...]	[1, 1, 1, ...]	1

Now the dataset is **ready to feed into DistilBERT**.