Popular NLP Models & Techniques with Examples

Natural Language Processing (NLP) has evolved significantly with advancements in machine learning and deep learning. Here's a breakdown of the most popular models and techniques used in NLP.

Swipe

1. Traditional NLP Techniques

(A) Bag-of-Words (BoW)

- Represents text as a collection of words without considering order.
- Used in text classification, spam filtering, and sentiment analysis.
- Creates a sparse matrix where each word's frequency is counted.

Example:

```python
from sklearn.feature_extraction.text import CountVectorizer

text = ["I love machine learning.", "Machine learning is amazing!"]
vectorizer = CountVectorizer()
bow_matrix = vectorizer.fit_transform(text)

print(vectorizer.get_feature_names_out())
print(bow_matrix.toarray())
```

Swipe

1. Traditional NLP Techniques

(B) TF-IDF (Term Frequency - Inverse Document Frequency)

- Improves BoW by assigning importance scores to words based on frequency.
- Common words like "the" and "is" get lower scores, while important words get higher scores.

Example:

```python
from sklearn.feature_extraction.text import TfidfVectorizer

text = ["I love NLP and deep learning.", "Deep learning is amazing!"]
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(text)

print(vectorizer.get_feature_names_out())
print(tfidf_matrix.toarray())
```

Swipe

1. Traditional NLP Techniques

(C) Word Embeddings (Word2Vec, GloVe, FastText)

- Represents words as dense vectors in a continuous vector space.
- Similar words have similar vector representations.
- Word2Vec (Google) & GloVe (Stanford) are widely used.

Example:

```python
from gensim.models import Word2Vec

sentences = [["I", "love", "NLP"], ["Deep", "learning", "is", "amazing"]]
model = Word2Vec(sentences, vector_size=10, min_count=1)

print("Vector for 'NLP':", model.wv["NLP"])
```

Swipe

2. Deep Learning-Based NLP Models

(A) Recurrent Neural Networks (RNNs)

- Good for sequential data like text but suffers from vanishing gradients.
- Used for text generation, machine translation, speech recognition.

(B) Long Short-Term Memory (LSTM)

- An improved RNN that remembers long-term dependencies.
- Used in chatbots, sentiment analysis, and speech recognition.

Example:

```python
from keras.models import Sequential
from keras.layers import LSTM, Dense, Embedding

model = Sequential([
    Embedding(input_dim=1000, output_dim=64),
    LSTM(128),
    Dense(1, activation='sigmoid')
])

model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
print(model.summary())
```

Swipe

(C) Transformers (BERT, GPT, T5, XLNet)

- Revolutionized NLP by handling long-range dependencies better than LSTMs.
- Self-attention mechanism allows parallel processing of words.
- Used for chatbots, search engines, and summarization.

(1) BERT (Bidirectional Encoder Representations from Transformers)

- Understands the context from both left and right sides of a word.
- Best for text classification, named entity recognition, and QA systems.

Example (Using Hugging Face BERT for Text Classification):

```
from transformers import pipeline

classifier = pipeline("text-classification", model="bert-base-uncased")

result = classifier("I love deep learning and NLP!")
print(result)
```

(3) T5 (Text-to-Text Transfer Transformer)

- Converts every NLP task into a text generation task.
- Used for translation, summarization, and QA.

Example (Using T5 for Summarization):

```python
from transformers import pipeline

summarizer = pipeline("summarization", model="t5-small")

text = "Natural Language Processing (NLP) is a field of AI that helps computers understand human language."

summary = summarizer(text, max_length=30, min_length=10, do_sample=False)
print("Summary:", summary[0]['summary_text'])
```

Swipe

3. Large Language Models (LLMs)

- GPT-4, LLaMA, PaLM, Claude
- These models learn from billions of words and perform complex reasoning.
- Used in AI-powered chatbots, code generation, and creative writing.

Choosing the Right Model

Task	Best Model
Sentiment Analysis	BERT, LSTM
Named Entity Recognition	BERT, SpaCy
Machine Translation	T5, MarianMT
Question Answering	BERT, T5
Speech Recognition	Whisper, DeepSpeech
Text Generation	GPT-4, GPT-2
Summarization	T5, Pegasus

Swipe

Conclusion

- Each NLP model has its strengths and is suited for different tasks.
- Are you planning to use any of these models for your AI chatbot project or Intelligent Document Processing (IDP) system?

@Birendra kumar