Table: Advantages and Disadvantages of One-Hot Encoding, Bag-of-Words (BoW), N-Grams, and TF-IDF

Method	Advantages	Disadvantages
One-Hot Encoding	- Simple and easy to understand.	- High-dimensional representation, leading to the curse of dimensionality.
	- Preserves the categorical nature of data.	- Does not capture semantic relationships between categories.
	- Useful for algorithms that require numerical input.	- Inefficient for large categorical variables.
	- Does not require scaling of features.	- Cannot handle out-of-vocabulary terms.
	- Can handle missing values effectively by representing them as all-zero vectors.	- Increases computational complexity for large datasets.
Bag-of-Words (BoW)	- Simple and straightforward representation of text data.	- Ignore word order and grammar, leading to loss of context.
	- Efficient for large datasets.	- Treats all words as equally important, ignoring word semantics.
	- Can be easily interpreted and visualized.	- Requires extensive preprocessing for better results.
	- Captures basic word frequency information.	- Doesn't handle typos or misspellings well.

	- Can be combined with various algorithms for text classification and clustering.	- Sparse representation can be memory-intensive for large vocabularies.
N-Grams	- Captures local word order information.	- Increases feature dimensionality, especially with higher n-values.
	- Preserves some context around words.	- Limited context windows may fail to capture long-range dependencies.
	- Useful for capturing short phrases and idiomatic expressions.	- Sensitive to noise and variability in language usage.
	- Can be used to model syntactic and semantic relationships.	- Requires careful selection of n-value.
	- Provides a compromise between Bag-of-Words and full sequence models.	- Memory and computation-intensive for large n-values and datasets.
TF-IDF	- Highlight important terms while downweighting common terms.	- Does not capture semantic relationships between words.
	- Effective in reducing the impact of noise words (stop-words).	- Requires a representative corpus for accurate IDF calculation.
	- Handles out-of-vocabulary terms gracefully.	- Sensitive to term variations and synonyms.
	- Captures the importance of a term within a document and across a corpus.	- May not perform well with short documents or small corpora.

	- Widely used in information retrieval, text mining, and document classification tasks.	- Computational overhead in calculating TF-IDF scores for large datasets.