NLP : Recurrent Neural Network

Video 5: Word Embedding

Analytics Vidhya

Analytics Vidhya

Beyond Basic Vectorization Techniques

Count Vectorization & TF-IDF

Bag of words Models

Sequence based Models

Vectorization

One Hot Encoding

Countvectorizer

TF-IDF

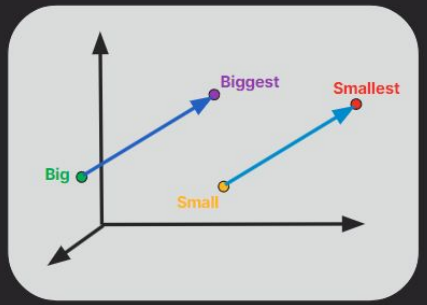
Word Embedding

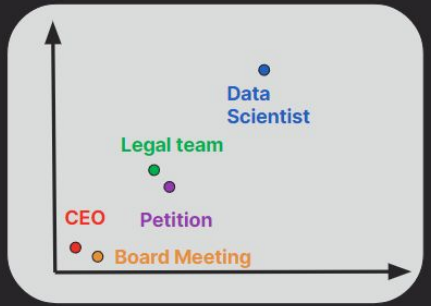
Word Embedding

Represents **words as vectors** of dimensions in a way that **captures their meanings, relationships, and context**.

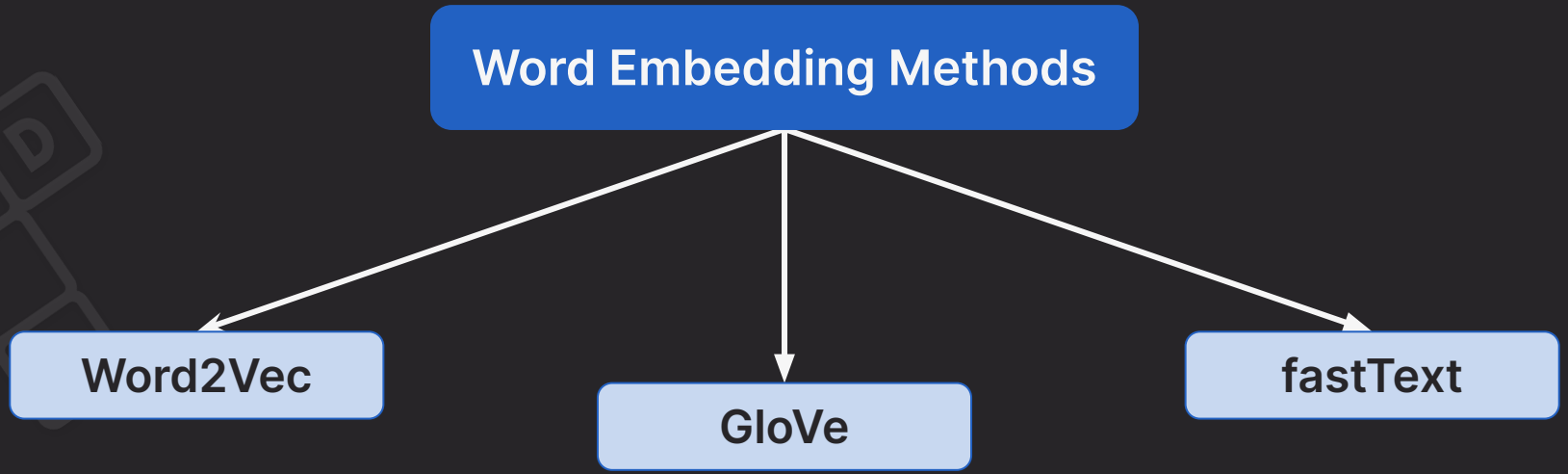
Vectors

CEO = [0.2, -0.7,.....,0.6]
CTO = [0.5, -0.6,.....,0.7]

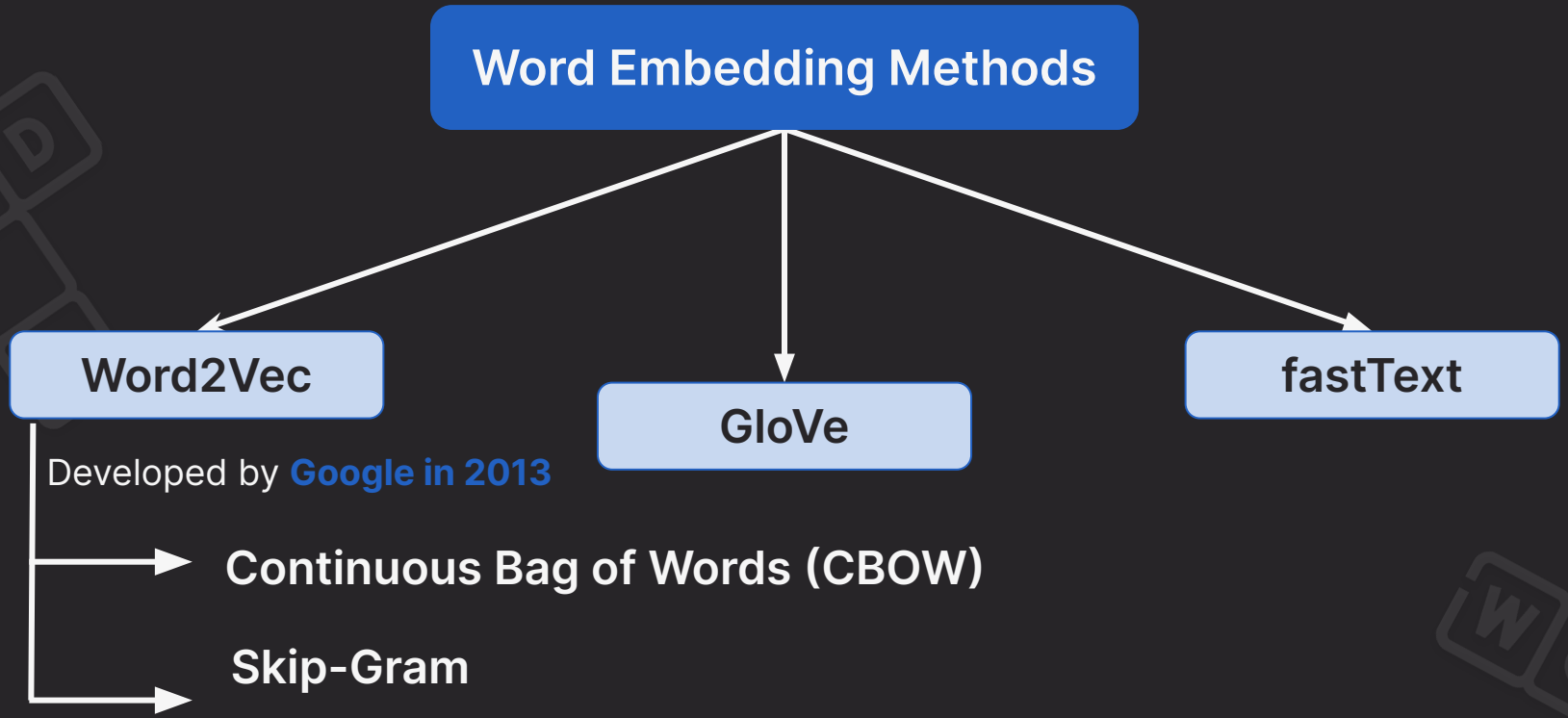




Word Embedding



Word Embedding



Analytics Vidhya

Continuous Bag of Words

"The CEO delivered a compelling presentation at the board meeting."

Analytics Vidhya

Continuous Bag of Words

Takes **context words as input** and predicts the most likely outcome or word to complete the missing part of sentence.

"The CEO **delivered** a **compelling** _____ at the **board meeting**."

presentation

speech

idea

thought

Analytics Vidhya

Continuous Bag of Words

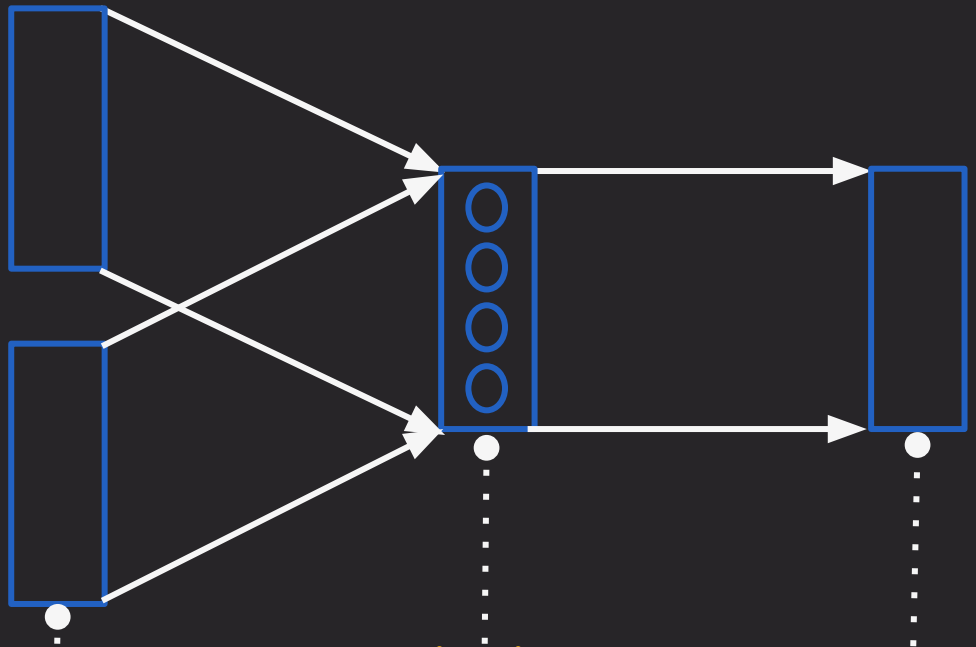
"The CEO **delivered** a **compelling** _____ at the **board meeting**."

Input	Target
delivered	presentation
compelling	presentation
board meeting	presentation

presentation speech idea thought

Analytics Vidhya

Continuous Bag of Words Model



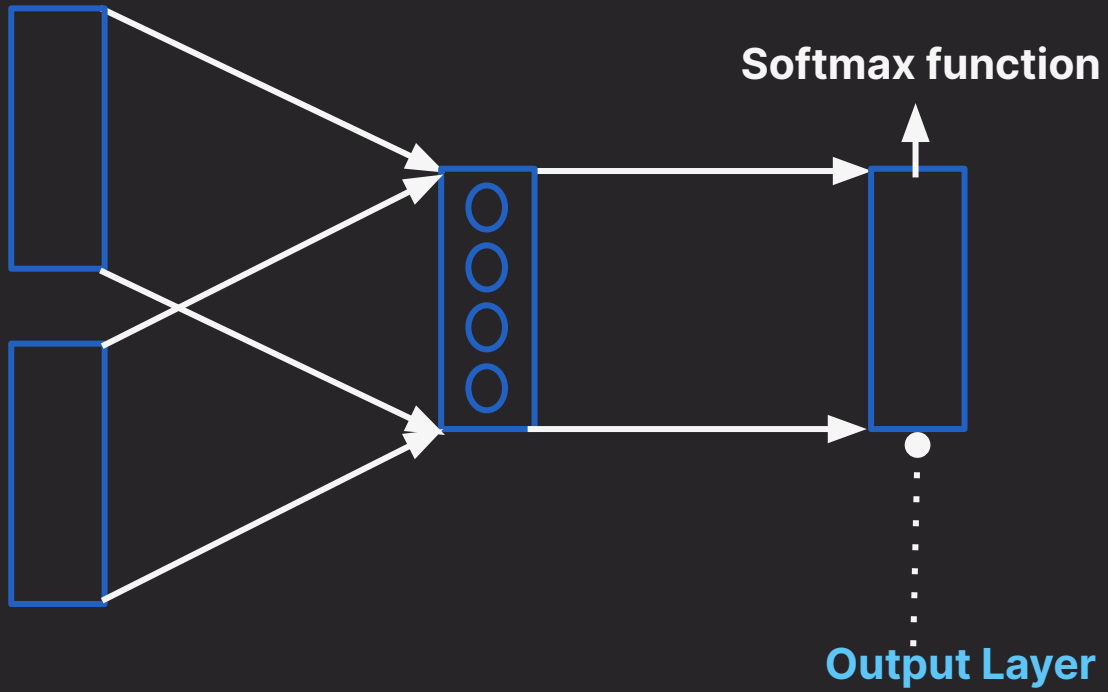
Input Layer
Context words

Projection Layer
Word embeddings

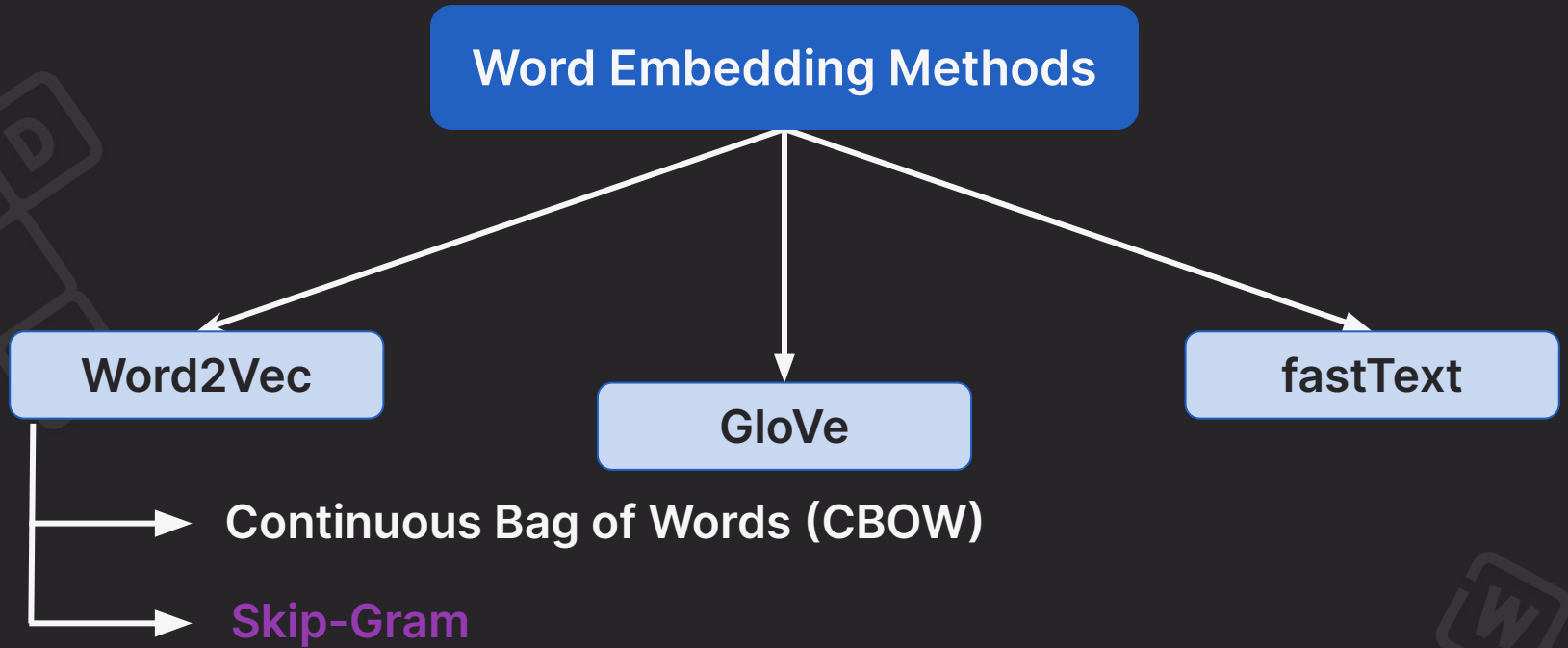
Output Layer
Predict the target word

Analytics Vidhya

Continuous Bag of Words Model



Word Embedding



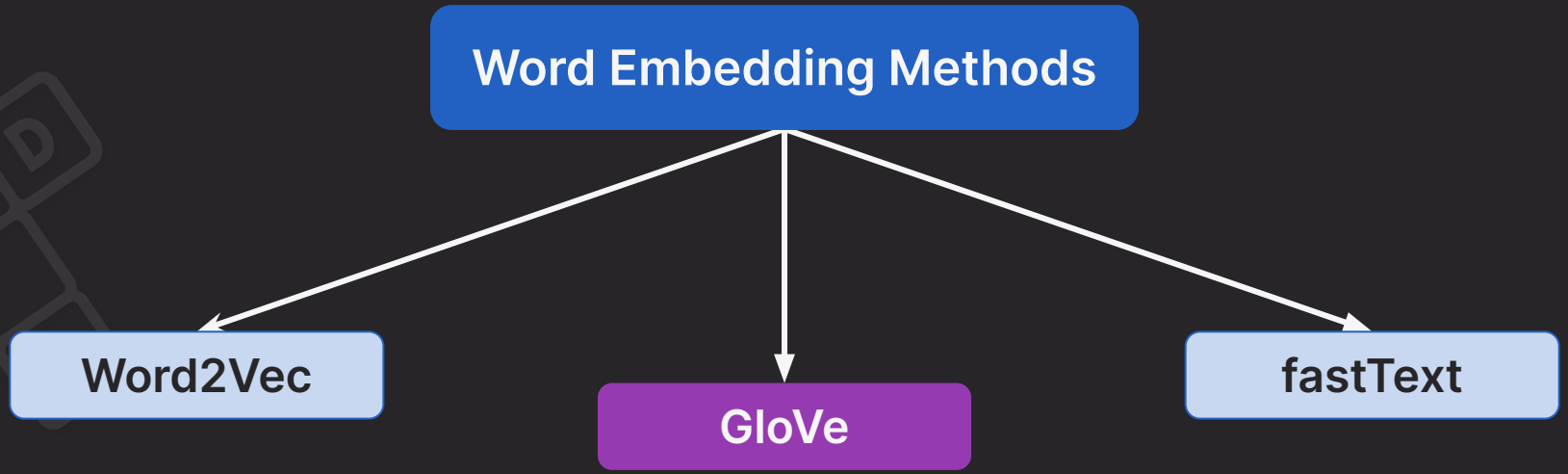
Skip-Gram

Starts with a target word and uses it to predict the surrounding context words.

"The CEO delivered a compelling **presentation** at the board meeting."

Input	Target
presentation	delivered
presentation	compelling
presentation	board meeting

Word Embedding



Analytics Vidhya

2. GloVe

- Stands for Global Vectors.
- Developed by **Stanford University researchers**.
- Analyzes frequency of words and assigns similar numerical vectors.

2. GloVe : Building the Co-occurrence Matrix

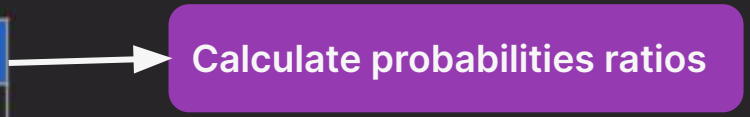
"The CEO delivered compelling presentation at the board meeting."

Window size = 1

	CEO	delivered	compelling	presentation	board
CEO	0	1	0	0	0
delivered	1	0	1	0	0
compelling	0	1	0	1	0
presentation	0	0	1	0	0
board	0	0	0	0	0

2. GloVe

	CEO	delivered	compelling	presentation	board
CEO	0	1	0	0	0
delivered	1	0	1	0	0
compelling	0	1	0	1	0
presentation	0	0	1	0	0
board	0	0	0	0	0



2. GloVe

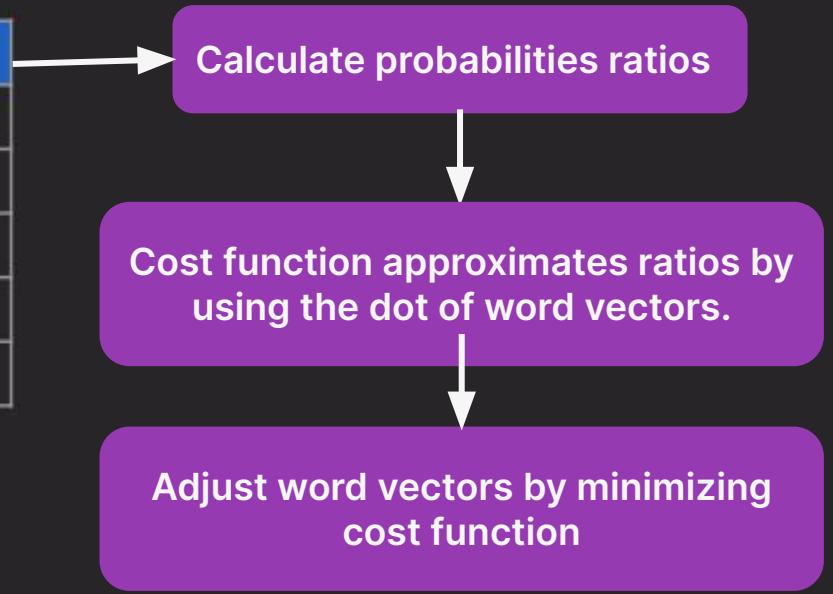
	CEO	delivered	compelling	presentation	board
CEO	0	1	0	0	0
delivered	1	0	1	0	0
compelling	0	1	0	1	0
presentation	0	0	1	0	0
board	0	0	0	0	0

Calculate probabilities ratios

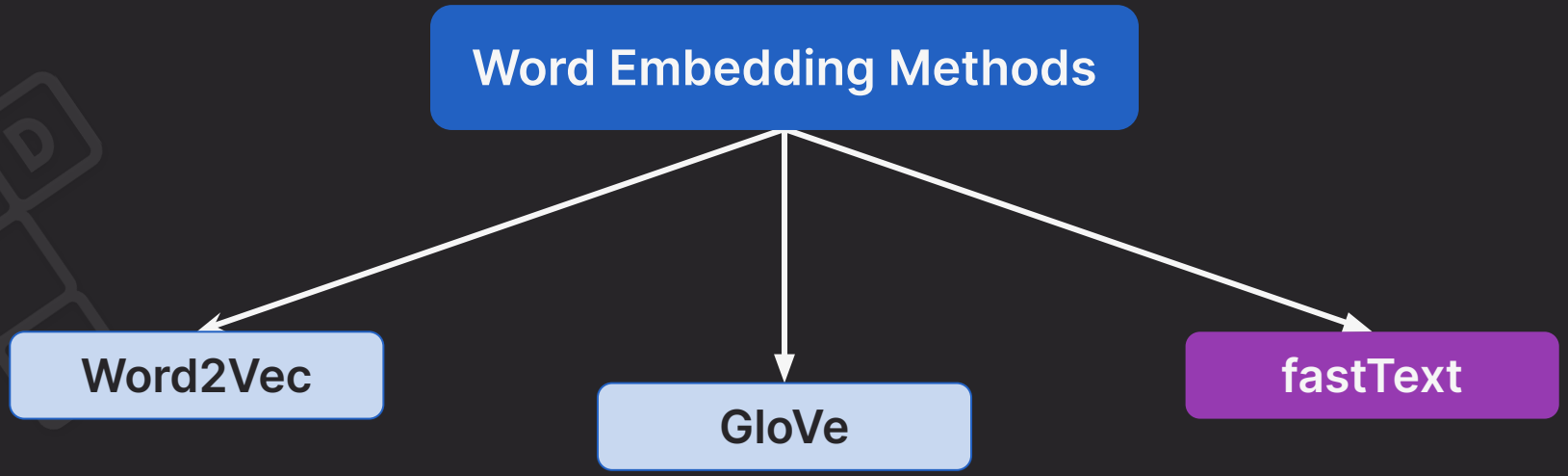
Cost function approximates ratios by using the dot of word vectors.

2. GloVe

	CEO	delivered	compelling	presentation	board
CEO	0	1	0	0	0
delivered	1	0	1	0	0
compelling	0	1	0	1	0
presentation	0	0	1	0	0
board	0	0	0	0	0



Word Embedding

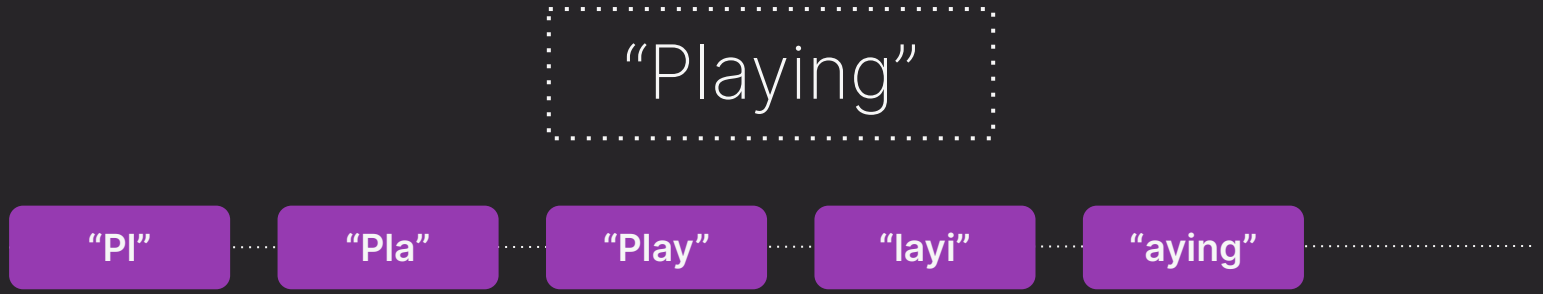


Analytics Vidhya

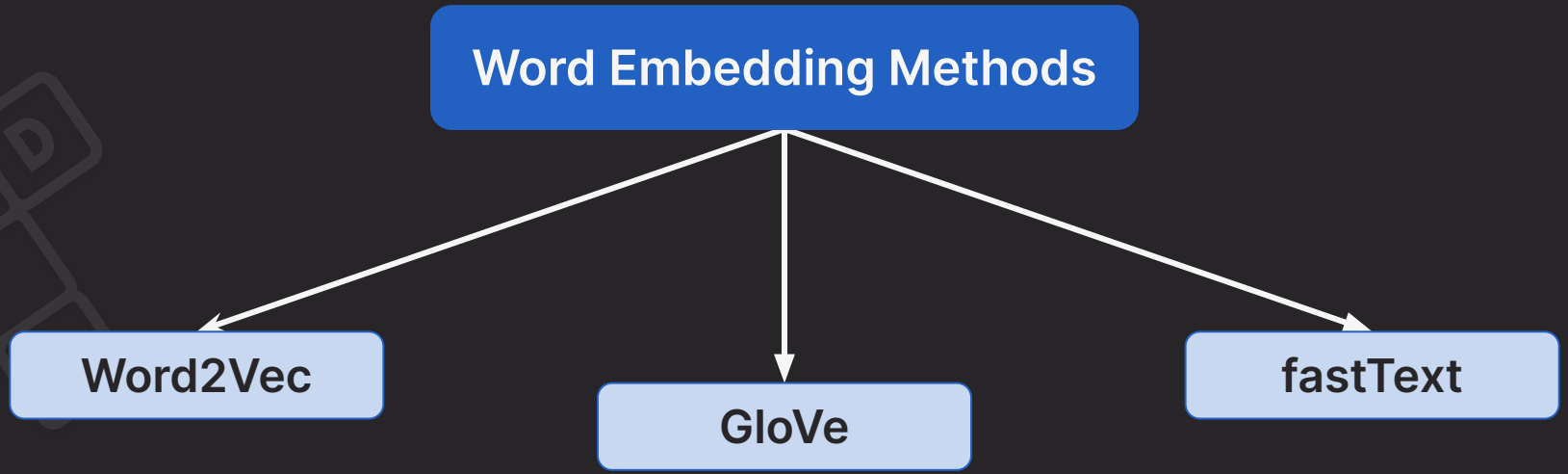
3. fastText

Developed by **Google in 2015**

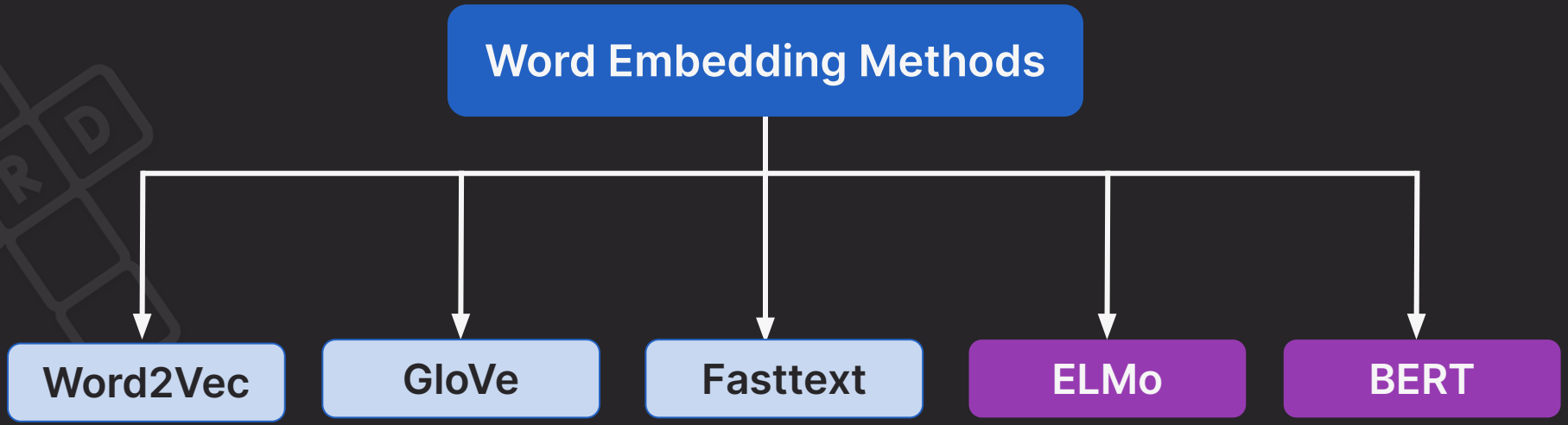
- Enhances word embeddings by using sub-word units like n-grams of characters within words.
- Extends the Word2Vec model to handle rare, new, or misspelled words.



Word Embedding



Word Embedding



4. ELMo

Embedding from language models

- Uses bi-directional LSTM models
- Pre-trained on large corpus to predict next word based on previous and future words.

Analytics Vidhya

5. BERT

Bidirectional Encoder Representations from Transformers

- Consider both left and right context of words in a sentence.
- Uses Transformer architecture

Analytics Vidhya

Up Next: RNN Model in Jupyter