### **Sites Referenced -** <https://www.algolia.com/blog/ai/vector-vs-keyword-search-why-you-should-care/>

<https://www.glean.com/blog/guide-to-vector-search>

<https://research.google/blog/open-sourcing-bert-state-of-the-art-pre-training-for-natural-language-processing/>

Step-by-step BERT semantic implementation - <https://myscale.com/blog/implementing-semantic-search-python-bert-step-by-step-guide/#>

### **Traditional Information Retrieval Techniques**

**1. Boolean Retrieval**

* **Pros**:
  + **Simplicity**: Easy to implement and understand.
  + **Performance**: Effective for specific queries with exact terms.
* **Cons**:
  + **Inflexibility**: Fails to understand synonyms or variations in phrasing.
  + **Overwhelming Results**: Can return too many or too few results.
* **Limitations**:
  + It doesn't handle ambiguity well, meaning user satisfaction is low when queries are vague​

**2. Vector Space Model (VSM)**

* **Pros**:
  + **Granular Ranking**: Provides ranked results based on cosine similarity.
  + **Better Flexibility**: Goes beyond exact matches, capturing some relationships between query terms.
* **Cons**:
  + **Manual Weighting**: Requires careful tuning of term weights (like TF-IDF).
  + **Resource-Intensive**: Computing cosine similarity for large datasets can be slow.
* **Limitations**:
  + Despite improved relevance over Boolean search, VSM struggles with synonyms and deeper semantic relationships​.

**3. BM25**

* **Pros**:
  + **Best in Class**: Considered one of the best-performing traditional IR models.
  + **Customizable**: Weights terms more intelligently than simple TF-IDF, improving relevance.
* **Cons**:
  + **Complexity**: More computationally expensive compared to simpler models like Boolean or VSM.
* **Limitations**:
  + BM25 is still limited in understanding deeper contextual and semantic nuances​

### **Modern Information Retrieval Techniques**

**1. Latent Semantic Analysis (LSA)**

* **Pros**:
  + **Captures Synonyms**: Identifies hidden relationships and synonymy between terms.
  + **Dimension Reduction**: Improves search by reducing term-dimensional space.
* **Cons**:
  + **Overfitting Risk**: Can misinterpret relationships if not properly tuned.
  + **Interpretation**: The reduced dimensional space is less interpretable than traditional methods.
* **Limitations**:
  + Requires significant computational power for larger datasets and lacks real-time performance in high-volume systems​

**2. Neural Information Retrieval (NIR)**

* **Pros**:
  + **Context-Aware**: Captures the meaning and context behind words, not just matching keywords.
  + **High Accuracy**: Delivers highly relevant results for complex, natural-language queries.
* **Cons**:
  + **Resource-Heavy**: Requires massive computational resources for training and querying.
  + **Data-Dependent**: Performance is heavily dependent on high-quality, labeled training data.
* **Limitations**:
  + Neural models are costly to train and maintain, and model updates may be slow depending on computational resources​

**3. Semantic Search**

* **Pros**:
  + **Synonym Handling**: Identifies intent and relations between query and document terms, not just word matches.
  + **Domain Flexibility**: Can be applied in various domains like healthcare and e-commerce, making it versatile.
* **Cons**:
  + **Complex Query Processing**: Requires more sophisticated algorithms, which may slow down response times.
* **Limitations**:
  + Real-time performance may suffer for large-scale applications without adequate hardware​

### **Deep Learning in IR**

**1. Sentence Transformers**

* **Pros**:
  + **High Precision**: Captures fine-grained semantic meaning for improved accuracy.
  + **Cross-Lingual Search**: Useful in multilingual applications, supporting queries in one language to retrieve documents in another.
* **Cons**:
  + **Infrastructure Heavy**: Requires vector databases to support embeddings, adding complexity.
* **Limitations**:
  + Scalability challenges arise as more embeddings need to be generated and stored for large datasets​

**Considered Industry best - Semantic Search using BERT transformers:**

**Pros** of BERT-based semantic search:

* **Contextual understanding**: It captures the nuances of language, providing highly relevant search results.
* **High accuracy**: Especially effective for complex queries and understanding user intent.
* **Flexibility**: Can be applied across various domains, including e-commerce, healthcare, and digital libraries.

**Cons**:

* **High computational cost**: Requires significant resources for training and deploying models.
* **Complex implementation**: Deep learning-based models are harder to implement compared to traditional techniques like TF-IDF or BM25.