Work flow ->

Pre-processing with n-gram

* Detect n-grams (using statistical measures like, PMI (Point-wise mutual information), chi-square)
* Phrase detection tools
* Update vocabulary
* Run LDA

1. Hybrid Approach

* Represent documents as a mix of unigram and n-gram

Deepseek conversation:

<https://chat.deepseek.com/a/chat/s/b712dee6-b638-4296-bde3-67ff0f21b8b1>

another one:   
<https://chat.deepseek.com/a/chat/s/8b3957fe-2613-43f6-87d4-9f8caa5a159a>

* Aggressive stemming use korte hbe.
* Next meeting e code base kaj dekhate hbe.
* 1st implement LDA

[Raw Corpus]

↓

[Text Cleaning + Stemming]

↓

[Tokenization + Bigram/Trigram Detection]

↓

┌───────────── ─┬────────────────────┐

│ │ │

↓ ↓ ↓

[N-gram Model] [Phrase-Merged Corpus] → [LDA Topic Model]

↓ ↓ ↓ ↓

[Top-k Continuations] ← Match Topic Context ←

↓

[Ranked Trend Predictions]