

Multilingual NLP Chatbot Project Report

1. Data Sources and Multilingual Composition

We curated documents in two languages: **English and Urdu**, centered on foundational topics in Artificial Intelligence (AI) and Machine Learning (ML).

- **English data** was sourced from publicly available educational material.
- **Urdu data** was gathered through manual translation and online Urdu-language resources.

Language	No. of Documents	Approx. Chunks After Processing
English	13	~150
Urdu	15	~170

2. Chunking Strategy and Preprocessing

Preprocessing:

- Stripped excess whitespaces and newlines.
- Standardized Unicode formatting (especially for Urdu).
- Lowercased English texts where applicable.

Chunking Strategy:

Strategy	Details
Fixed-Length Chunking	Each document was split into 100-token chunks with 20-token overlap
Metadata	Each chunk stored with <code>filename</code> , <code>language</code> , and <code>chunk_id</code>

This ensures semantic coherence and improves retrieval quality during search.

3. Embedding Models and Vector Store Architecture

We used **three sentence embedding models** to encode the English and Urdu document chunks into high-dimensional vector space.

Model Name	Type	Embedding Size	Language Support

DistilUSE (distiluse-base-multilingual-cased)	Multilingual	512	English + Urdu
BGE-small (BAAI/bge-small-en-v1.5)	English-only	384	English only
E5-small (intfloat/multilingual-e5-small)	Multilingual	384	English + Urdu

Embeddings were stored using the .npy format for both languages separately and then uploaded to **Pinecone** for similarity search.

Vector Store (Pinecone):

- Separate indexes created per model:

multilingual-nlp-distiluse, multilingual-nlp-bge, multilingual-nlp-e5

- Metadata stored per vector (language, filename, chunk number)
- Search performed using **cosine similarity**

4. Results & Screenshots

English Query: "What is the difference between AI and Machine Learning?"

Model	Top Match Summary	Score
DistilUSE	ML is a subset of AI...	0.56
E5-small	They are not the same... relationship between AI and ML...	0.91
BGE-small	ML enables machines to learn from data...	0.86

Urdu Query: "مصنوعی ذہانت اور مینیٹ لرننگ کے درمیان کیا فرق ہے؟"

Model	Top Match Summary (Translated)	Score
DistilUSE	مصنوعی ذہانت ایک مکمل شعبہ ہے...	0.27
E5-small	مینیٹ لرننگ کو مصنوعی ذہانت کا حصہ سمجھا جاتا ہے...	0.91
BGE-small	مینیٹ لرننگ موجودہ دور میں ایک ایم موضع ہے...	0.91

 **Screenshots** of the working query system have been attached:

- 1.jpg:

```
1s 1 query_all_models(queries["english"])

Query: What is the difference between AI and machine learning?

--- Results from multilingual-nlp-distiluse ---
[english] (0.5635) → what is machine learning? machine learning (ml) is a subset of artificial intelligence that enables machines to learn from data without being
[english] (0.5243) → they are not the same thing but are closely connected. relationship between ai and ml understanding the relationship between ai and ml is imp
[english] (0.4842) → variety of approaches and algorithms, and machine learning being one of it.

--- Results from multilingual-nlp-e5 ---
[english] (0.9178) → they are not the same thing but are closely connected. relationship between ai and ml understanding the relationship between ai and ml is imp
[english] (0.9112) → type of learning, the machine is given unlabeled datasets to algorithms to find hidden patterns or data groupings. there are three types of i
[english] (0.9104) → what is machine learning? machine learning (ml) is a subset of artificial intelligence that enables machines to learn from data without being

--- Results from multilingual-nlp-bge ---
[english] (0.8594) → they are not the same thing but are closely connected. relationship between ai and ml understanding the relationship between ai and ml is imp
[english] (0.8127) → what is machine learning? machine learning (ml) is a subset of artificial intelligence that enables machines to learn from data without being
[english] (0.7973) → type of learning, the machine is given unlabeled datasets to algorithms to find hidden patterns or data groupings. there are three types of i
```

- 2.jpg:

```
1s 1 query_all_models(queries["urdu"])

Query: اور متنی لرنگ کے درمیان کی فرق ہے؟

--- Results from multilingual-nlp-distiluse ---
[urdu] (0.2721) → ان متصدی کے حصول کے لیے خاص شاخ ہے جو متنوں کو خوبصورت سیکھ کر قابل بذائقے کرے
[urdu] (0.2461) → ریاضیاتی پہلو، مطلوب اصول، مصنوعی اصلاحی جد، اور شماریاتی صلاحیت تھی، اور
[urdu] (0.2375) → اپنی الفاظ میں، متن مطابق میں "بیکاری" اور پھر اسی بیکاری کی بذریعہ برداری لگاتی ہے اور فصلیت کرتی ہے۔ وہ گزولی کے سبق جب متن کو مزید مطابق دی جاتی ہے اور
[urdu] (0.9117) → بیکاری اور مسجدی بوجہ حاصل کرنے کی مصالحت ہوتی ہے۔ زیاد، مصنوعی نتائج اور فصلیت کرتی ہے اور فصلیت کرتی ہے وہ گزولی کے سبق جب متن کو زیر مطابق دی جاتی ہے اور
[urdu] (0.8962) → اپنی الفاظ میں، متن مطابق میں "بیکاری" اور پھر اسی بیکاری کی بذریعہ برداری لگاتی ہے اور فصلیت کرتی ہے وہ گزولی کے سبق جب متن کو زیر مطابق دی جاتی ہے اور
[urdu] (0.8907) → اپنی الفاظ میں، مصنوعی نتائج اور اسکالاٹ اکثر ایک جوئی مسجدی جاتی ہے، مگر ان میں فرق ہے: مصنوعی نتائج ایک اور مسلم صورتی ہے جس کو انسان دے
[urdu] (0.9101) → جید تخلیقی مصنوعی نتائج کی آمد، اور ان کی مصالحت کہ وہ نیا مولدا اور بیتلہ کر سکے، نے ملکی غیر منافق نسلات اور خطرات کو اسے تلب کیا این نے موجودہ اور مستقبل میں مصنوعی نتائج کے
[urdu] (0.9092) → رد مقصودی سے اپسی اقسام کریں کہ فلک دالیے ہیں جو متنیں کردہ مقاصد کو حاصل کریں کے امکالت کو دیتا ہے اور اسی متنوں کو مصنوعی نتائج رکھنے والی متنیں کیا جاتی ہے مصنوعی نتائج کی
[urdu] (0.9047) → ای مصنوعی نتائج کپلا کر تپڑی بچھوپی جاتی، کیونکہ جب کوئی جزو علم اور مفہود یو جلتے تو اسے مصنوعی نتائج کہنا جوڑتا ہے۔ مصنوعی نتائج کی تحقیق مختلف مقاصد اور مخصوص اور زاویوں کے
```

5. Comparative Table – Model Evaluation

Model	Language Support	Embedding Size	Avg. Score	Strengths	Limitations
DistilUSE	English + Urdu	512	~0.50	Lightweight, multilingual	Slightly lower accuracy
BGE-small	English-only	384	~0.85	Strong contextual accuracy	Not suitable for Urdu
E5-small	English + Urdu	384	~0.91	Best multilingual performance	Slightly heavier computation

Conclusion

This project successfully implemented a multilingual NLP vector search system using:

- Parallel English and Urdu corpora
- Three modern embedding models
- Chunking and preprocessing techniques
- Pinecone vector database for fast semantic search

