Code: 20AM3603

III B.Tech - II Semester – Regular Examinations - APRIL 2025

NATURAL LANGUAGE PROCESSING
(ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

Duration: 3 hours   Max. Marks: 70

Note: 1. This paper contains questions from 5 units of Syllabus. Each unit carries 14 marks and have an internal choice of Questions.
2. All parts of Question must be answered in one place.

BL – Blooms Level   CO – Course Outcome

			BL	CO	Max. Marks
		UNIT-I			
1	a)	Explain the differences between syntax, semantics,and pragmatics in the context of language understanding.	CO1	L2	7 M
	b)	Discuss the primary applications of NLP in daily life. Provide at least three examples.	CO1	L2	7 M
		OR			
2	a)	How can text be represented as words, sentences and documents in NLP?	CO1	L2	7 M
	b)	Describe tokenization, stemming and lemmatization? Provide examples for each.	CO1	L2	7 M
		UNIT-II			
3	a)	Explain the process of removing stop words, punctuation and special characters from text.	CO2	L3	7 M

	b)	Illustrate text normalization and how it handles misspellings and variations in text.	CO2	L3	7 M
OR					
4	a)	Explain N-grams and their role in NLP. How do they help in feature extraction?	CO2	L3	7 M
	b)	How can text features be evaluated and visualized? Describe at least two techniques.	CO2	L3	7 M
UNIT-III					
5	a)	Compare BERT with traditional language models in NLP.	CO3	L3	7 M
	b)	How can BERT be fine-tuned for specific downstream NLP tasks?	CO3	L3	7 M
OR					
6	a)	Explain real-world applications of BERT in natural language understanding.	CO4	L4	7 M
	b)	Illustrate the advantages of transfer learning in BERT for NLP applications.	CO3	L3	7 M
UNIT-IV					
7	a)	Compare the key differences between traditional language models and Transformer-based models?	CO4	L4	7 M
	b)	Explain multi-head attention and how does it enhance the representation diversity in Transformers?	CO3	L3	7 M
OR					
8	a)	Explain the role of Transformers in language summarization and modeling.	CO3	L3	7 M

	b)	Explain real-world examples of a Transformer-based model.	CO4	L4	7 M
UNIT-V					
9	a)	Compare and Contrast GPT and BERT models.	CO4	L4	7 M
	b)	Analyse the impact of biases on GPT and its output.	CO4	L4	7 M
OR					
10	a)	Illustrate the ethical consideration of large language models like GPT.	CO3	L3	7 M
	b)	Explain GPT model and anlayse its performance in program generation and code completion tasks.	CO3	L3	7 M

PVP20

Code No:20AM3603

III B. Tech – II Semester-Regular Examinations-APRIL2025
NATURAL LANGUAGE PROCESSING (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

Duration: 3 Hours **Max. Marks: 70**

Note:

1. This question paper contains questions from 5 units of Syllabus. Each Unit carries 14 Marks and have an internal choice of Questions.
2. All parts of Question paper must be answered in one place.

5 x 14 = 70 Marks

1(a) Explain the differences between syntax, semantics, and pragmatics in the context of language understanding. ***(7M)***

- **Syntax:** Structure or grammar of a sentence. [3Marks]
- **Semantics:** Meaning of words and sentences. [2Marks]
- **Pragmatics:** Meaning based on context. [2Marks]

1(b) Discuss the primary applications of NLP in daily life. Provide at least three examples. ***(7M)***

2-3 Marks for each application

2(a) How can text be represented as words, sentences, and documents in NLP? ***(7M)***

1. **Words** – "NLP is fun" = [NLP, is, fun] (3Mark)
2. **Sentences** – Whole sentences used as input.(2Mark)
3. **Documents** – Large texts used in training or analysis.(2Marks)

2(b) Describe tokenization, stemming, and lemmatization. Provide examples for each. ***(7M)***

- **Tokenization:** Splitting text into words. "I love cats" -> [I, love, cats] [3Marks]
- **Stemming:** Removing endings. "Playing" -> "play" [2Marks]
- **Lemmatization:** Getting base word. "Better" -> "good" [2Marks]

3(a)Explain the process of removing stop words, punctuation, and special characters from text. ***(7M)***

- Stop word removal **(2 marks)**
- Punctuation removal **(2 marks)**
- Special character removal **(2 marks)**
- Purpose/importance of cleaning **(1 mark)**

3(b) Illustrate text normalization and how it handles misspellings and variations in text. ***(7M)***

- Lowercasing **(2 mark)**
- Removing misspellings/contractions **(3 marks)**
- Handling variations in formats (dates, numbers) **(2 marks)**

4(a) Explain N-grams and their role in NLP. How do they help in feature extraction? ***(7M)***

- Definition of N-gram and it's types **(3 marks)**
- Example (unigram, bigram, trigram) **(2 marks)**
- Helps model understand word combinations. **(2 marks)**

4(b) How can text features be evaluated and visualized? Describe at least two techniques. ***(7M)***

- Visualizing embeddings using t-SNE and PCA **(5 marks)**
- Application in understanding text features **(2 marks)**

5(a) Compare BERT with traditional language models in NLP. ***(7M)***

- Bidirectional vs unidirectional **(2 marks)**
- Context understanding/comparison **(2 marks)**
- Speed **(1 mark)**
- Task performance (fine-tuning on multiple tasks) **(2 marks)**

5(b) How can BERT be fine-tuned for specific downstream NLP tasks? ***(7M)***

- Load pretrained BERT model **(4 marks)**
- Add task-specific output layer **(3 marks)**

6(a) Explain real-world applications of BERT in natural language understanding. ***(7M)***

2-3 Marks for each application

6(b) Illustrate the advantages of transfer learning in BERT for NLP applications. ***(7M)***

- First train on big data. **(3 marks)**
- Then re-use it for smaller tasks. **(2 marks)**
- Saves time and gives better results. **(2 marks)**

7(a) Compare the key differences between traditional language models and Transformer-based models. ***(7M)***

- Unidirectional vs. Bidirectional/Autoregressive **(2 marks)**
- Sequential processing vs. Parallel attention **(2 marks)**
- Applications**(2 marks)**

- Additional valid comparison or example **(1 marks)**

7(b) Explain multi-head attention and how does it enhance the representation diversity in Transformers? *(7M)*

- Definition of Multi-head Attention (parallel attention heads) **(3marks)**
- Role of different heads capturing different relationships **(2 marks)**
- How it enhances diversity in representations **(2 marks)**

8(a) Explain the role of Transformers in language summarization and modeling. *(7M)*

- Use of encoder-decoder for summarization **(3 marks)**
- Advantages: long context, bidirectional context **(2 marks)**
- Real-world applications (e.g., summarization, text gen.) **(2 mark)**

8(b) Explain real-world examples of Transformer-based model. *(7M)*

2-3 Marks for each application

9(a) Compare and Contrast GPT and BERT models. *(7M)*

- GPT is unidirectional; BERT is bidirectional **(2 marks)**
- GPT for generation; BERT for classification **(2 marks)**
- Real-world usage or concluding summary **(2 marks)**
- GPT predicts next word; BERT uses MLM **(1 marks)**

9(b) Analyse the impact of biases on GPT and its output. *(7M)*

- Training Data Bias (e.g., stereotypes) **(3 marks)**
- Language or Decision-Making Bias **(2 marks)**
- Ethical concerns from biased outputs **(2 marks)**

10(a) Illustrate the ethical consideration of large language models like GPT. *(7M)*

- May give wrong or fake info. **(2 marks)**
- May remember private data. **(2 marks)**
- Can be used to create spam. **(2 marks)**
- Uses a lot of electricity to train. **(1 marks)**

10(b) Explain GPT model and anlayse its performance in program generation and code completion tasks. *(7M)*

- GPT architecture overview **(3 mark)**
- Pretraining and fine-tuning steps **(2 marks)**
- Strengths in code generation (syntax, basic tasks) **(2 marks)**

PVP20

Code No:20AM3603

III B. Tech – II Semester-Regular Examinations-APRIL2025
NATURAL LANGUAGE PROCESSING (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

Duration: 3 Hours **Max. Marks: 70**

Note:

1. This question paper contains questions from 5 units of Syllabus. Each Unit carries
   14 Marks and have an internal choice of Questions.
2. All parts of Question paper must be answered in one place.

5 x 14 = 70 Marks

1(a) Explain the differences between syntax, semantics, and pragmatics in the context of language understanding. ***(7M)***

Answer:

- **Syntax:**
  - Deals with the grammatical structure of sentences.
  - Concerned with how words are combined to form valid sentences.
- **Semantics:**
  - Focuses on the meaning of words and sentences.
  - Ensures that the sentence makes logical sense.
- **Pragmatics:**
  - Deals with the context and intended meaning behind the words.
  - Interprets how real-world context affects language understanding.

1(b) Discuss the primary applications of NLP in daily life. Provide at least three examples. ***(7M)***

Answer: (Any three)

1. **Virtual Assistants** – Siri, Alexa use NLP to understand voice commands.
2. **Machine Translation** – Google Translate converts text between languages.
3. **Chatbots** – Answer queries in customer service.
4. **Text Prediction/Autocorrect** – Suggests words and corrects spelling.
5. **Sentiment Analysis** – Detects emotions in reviews or social media.
6. **Email Filtering** – Filters spam using text analysis.
7. **Search Engines** – Understand user queries and rank results.
8. **Voice-to-Text** – Converts speech to written text.
9. **Text Summarization** – Creates short summaries of long texts.

10. **Grammar Checkers** – Tools like Grammarly correct grammar and style.

OR

2(a) How can text be represented as words, sentences, and documents in NLP? *(7M)*

Answer:

1. **Word Representation:**
   - Individual words are represented as tokens (e.g., "NLP" → `['NLP']`).
   - Can be further converted into vectors .
2. **Sentence Representation:**
   - A sequence of word tokens forms a sentence.
   - Often represented using sentence embeddings (e.g., using BERT).
3. **Document Representation:**
   - A document is a collection of sentences.
   - Represented using models like TF-IDF, Doc2Vec, or averaged sentence embeddings.

2(b) Describe tokenization, stemming, and lemmatization. Provide examples for each. *(7M)*

Answer:

- **Tokenization:**
  - Splitting text into smaller units (tokens).
  - Example: "NLP is fun" → `['NLP', 'is', 'fun']`.
- **Stemming:**
  - Reduces words to their root form by chopping suffixes.
  - Example: "running", "runner" → "run".
  - May not produce actual words.
- **Lemmatization:**
  - Reduces words to their dictionary form using linguistic knowledge.
  - Example: "was", "is" → "be"; "running" → "run".
  - More accurate than stemming.

UNIT - II

3(a) Explain the process of removing stop words, punctuation, and special characters from text. *(7M)*

Answer:

- **Stop Words Removal:**
  - Removes common words like "is", "the", "and" which add little meaning.
  - Improves model efficiency and focuses on meaningful content.
- **Punctuation Removal:**
  - Removes symbols like `.` `,` `!` `?` to clean text.
  - Helps prevent irrelevant tokens from affecting results.
- **Special Characters Removal:**
  - Eliminates characters like `#`, `@`, `&`, `%`, which are not useful in most cases.
  - Done using regular expressions or string filtering.

3(b) Illustrate text normalization and how it handles misspellings and variations in text. ***(7M)***

Answer:

1. **Text Normalization:**
   - Converts text into a standard format.
   - Essential for consistency in NLP tasks.
2. **Handling Misspellings:** (anyone)
   - Lowercasing: "Apple" → "apple"
   - Removing punctuation: "hello!" → "hello"
   - Expanding contractions: "don't" → "do not"
   - Spelling correction: "teh" → "the"
   - Slang handling: "u" → "you"
3. **Handling Variations:**
   - Normalizing dates, numbers, currency formats.
   - Example: "$5.00" → "five dollars"

OR

4(a) Explain N-grams and their role in NLP. How do they help in feature extraction? ***(7M)***

Answer:

- **N-gram Definition:**
  - A sequence of N words from a given text.
  - Types: Unigram (1-word), Bigram (2-word), Trigram (3-word), etc.
- **Example:**
  - Sentence: "Natural Language Processing"

    - Bigrams: ("Natural Language"), ("Language Processing")
- **Role in NLP:**
    - Captures word context and sequence.
    - Helps models understand word relationships.

4(b) How can text features be evaluated and visualized? Describe at least two techniques. ***(7M)***

Answer:

- **Word Embeddings Visualization:**
    - Tools like **t-SNE** or **PCA** reduce word vectors to 2D or 3D for visualization.
    - Shows semantic similarity (e.g., "king" close to "queen").
- **Feature Importance Charts:**
    - Plots showing the most influential words for model predictions.
- **Word Clouds:**
    - Visually displays most frequent words in a text.

UNIT - III

5(a) Compare BERT with traditional language models in NLP. ***(7M)***

Answer: (any 4 points)

Traditional Language Models	**BERT (Bidirectional Encoder Representations from Transformers)**
Unidirectional (usually left-to-right or right-to-left)	Bidirectional – considers both left and right context simultaneously
Based on N-gram models, RNNs, or LSTMs	Based on Transformer encoder architecture
Predict the next word in a sequence	Masked Language Modeling (MLM) + Next Sentence Prediction (NSP)
Limited understanding (depends on past or future only)	Deep understanding using full sentence context
Sequential word processing → **slower training**	Parallel word processing using attention → **faster training**
Needs to be trained from scratch or modified per task	Pretrained model can be **fine-tuned** for many downstream tasks
struggles with complex tasks	Achieves **state-of-the-art** results on various NLP benchmarks

5(b) How can BERT be fine-tuned for specific downstream NLP tasks? *(7M)*

Answer:

1. **Process:**
   - Load pre-trained BERT.
   - Add appropriate output layer.
   - Train on task-specific dataset (fine-tuning all layers or just top).
2. **Examples of Tasks:**
   - Sentiment analysis, Named Entity Recognition, Question Answering.
3. **Add Task-Specific Layer:**
   - Append a classification or regression head to BERT's output.
4. **Advantages:**
   - Requires less data and training time than training from scratch.
   - Leverages learned representations.

OR

6(a) Explain real-world applications of BERT in natural language understanding. *(7M)*

Answer: (any 4 points)

- **Question Answering** – Finds precise answers from text (e.g., Google Search, chatbots).
- **Sentiment Analysis** – Understands positive/negative tone in reviews or tweets.
- **Spam Detection** – Classifies emails based on intent and context.
- **Text Summarization** – Generates concise summaries of long content.
- **Named Entity Recognition (NER)** – Identifies names, places, dates in text.
- **Semantic Search** – Matches queries with meaning, not just keywords.
- **Text Classification** – Categorizes documents (e.g., News, Sports, Legal).

6(b) Illustrate the advantages of transfer learning in BERT for NLP applications. *(7M)*

Answer: (any 3 points)

1. **Pre-trained Knowledge:**
   - BERT is trained on massive datasets (e.g., Wikipedia).
2. **Reduces Data Requirement:**
   - Effective even with small labeled datasets.

3. **Improves Accuracy:**
    - Outperforms traditional models across benchmarks.
4. **Faster Development:**
    - Minimal changes needed to adapt to new tasks.
5. **Versatile Use:**
    - Can be fine-tuned for classification, NER, QA, etc.

UNIT - IV

7(a) Compare the key differences between traditional language models and Transformer-based models. *(7M)*

Answer: (any 3 points)

Traditional Language Models	**Transformer-based Models (e.g., BERT, GPT)**
Based on **N-grams, RNNs,** or **LSTMs.**	Based on the **Transformer** architecture (self-attention).
Often unidirectional (left-to-right or right-to-left).	Bidirectional or autoregressive (depends on the model type).
Processes text **sequentially,** one word at a time.	Processes **entire input in parallel** using attention mechanisms.
Limited context (usually considers only the previous words).	Considers **global context** (all words in a sentence at once).
Typically trained to **predict the next word** (e.g., language modeling).	Trained with **Masked Language Modeling (MLM)** or **Next Sentence Prediction (NSP)** (for BERT).
Slower **due to sequential processing of text.**	Faster **due to parallel processing with self-attention.**
Generally performs well on simpler tasks, but struggles with long-range dependencies.	Achieves **state-of-the-art** performance across a wide range of NLP tasks.
Usually trained for a **single task** from scratch.	Pretrained on large corpora and **fine-tuned** for specific tasks.

7(b) Explain multi-head attention and how does it enhance the representation diversity in Transformers? *(7M)*

Answer:

1. **Multi-head Attention:**
    - Applies attention multiple times in parallel with different learned projections.

- Each "head" focuses on different parts of the sequence.

2. **Enhancing Diversity:**
    - Captures multiple types of relationships (e.g., syntactic, semantic).
    - Improves model's ability to understand complex patterns.
3. **Example:**
    - One head might focus on verbs, another on nouns in a sentence.

OR

8(a) Explain the role of Transformers in language summarization and modeling. ***(7M)***

Answer:

1. **Language Summarization:**
    - Use encoder-decoder architecture.
2. **Language Modeling:**
    - Predict next word/token (GPT), or masked tokens (BERT).
    - Learn contextual word usage.
3. **Advantages:**
    - Better handling of long texts.
    - Supports bidirectional context understanding.
4. **Applications:**
    - News summarization, auto-email reply, text generation.
    -

8(b) Explain real-world examples of Transformer-based model. ***(7M)***

Answer: (any 3 points)

Transformer-based models like **BERT, GPT,** and **RoBERTa** have revolutionized NLP by enabling more accurate and contextually aware understanding of text. These models are widely applied in real-world tasks like:

- Search engines **(e.g., Google Search),**
- Customer service **(e.g., chatbots),**
- Content generation **(e.g., GPT-3),**
- Document processing **(e.g., summarization, classification),**
- Translation **and** speech recognition.

UNIT – V

9(a) Compare and Contrast GPT and BERT models. *(7M)*

Answer: (any 4 points)

GPT (Generative Pretrained Transformer)	**BERT (Bidirectional Encoder Representations from Transformers)**
Autoregressive (unidirectional), generates text.	**Bidirectional**, reads text in both directions (left-to-right and right-to-left).
Next word prediction (predicts the next word in sequence).	**Masked Language Model (MLM)** (predicts missing words in a sentence).
Designed for **text generation** (creating coherent sequences).	Designed for **understanding** and **classifying** text.
Unidirectional (left-to-right).	**Bidirectional** (considers both left and right context).
Fine-tuned for **specific tasks** like text generation.	Fine-tuned for a wide range of tasks (classification, QA, NER, etc.).
Used in **text generation, chatbots,** and **creative writing.**	Used in **question answering, sentiment analysis, NER,** etc.
Excels in **text generation** and **language modeling.**	Excels in **contextual understanding** and **task-specific NLP** tasks.

9(b) Analyse the impact of biases on GPT and its output. *(7M)*

Answer: (any 3 points)

Training Data Bias: GPT learns biases from its training data, which can include harmful stereotypes.

Language Bias: GPT can favor certain linguistic styles or language structures, excluding others.

Decision-Making Bias: Bias in GPT affects tasks like sentiment analysis and recommendations.

Mitigation Efforts: Developers use debiasing techniques, but some biases persist.

Ethical Concerns: Biased outputs can lead to unfair treatment, especially in sensitive fields.

OR

10(a) Illustrate the ethical consideration of large language models like GPT. *(7M)*

Answer: (any 3 points)

- **Bias and Fairness**: GPT may produce biased outputs, reinforcing harmful stereotypes.
- **Misinformation**: GPT can generate false or misleading information, spreading fake news.
- **Privacy**: Risk of revealing sensitive or private data unintentionally.
- **Accountability**: Unclear responsibility for harmful consequences of model outputs.
- **Job Displacement**: AI automation may lead to job loss in fields like content creation and customer service.
- **Lack of Transparency**: Difficulty in understanding model decisions, affecting trust.
- **Manipulation**: Potential for misuse in activities like scams or creating deepfakes.
- **Environmental Impact**: High energy consumption during model training contributes to carbon footprint.

10(b) Explain GPT model and anlayse its performance in program generation and code completion tasks. *(7M)*

Answer:

1. Architecture: GPT uses unidirectional (autoregressive) Transformer architecture.
2. Pretraining: Trained on large text corpora to learn language patterns.
3. Fine-tuning: Can be fine-tuned for specific tasks (e.g., code generation, text generation).

Performance:

1. Code Generation:
   - Strengths: Generates syntactically correct code snippets for simple tasks.
   - Weaknesses: Struggles with complex logic and produces suboptimal solutions.
2. Code Completion:
   - Strengths: Predicts next lines of code and completes common patterns quickly.
   - Weaknesses: May generate incorrect or inefficient code for complex scenarios.