SRN P E S 2 P 4 E 2 3 D S 0 45

PES	**PES University, Bengaluru** (Established under Karnataka Act No. 16 of 2013)
	UE20CS933

July 2024: END SEMESTER ASSESSMENT (ESA)
M TECH DATA SCIENCE AND MACHINE LEARNING_ SEMESTER II

UE20CS933 - NATURAL LANGUAGE PROCESSING

Time: 3 Hrs	Answer All Questions	Max Marks: 100

INSTRUCTIONS

- All questions are compulsory.
- Section A should be handwritten in the answer script provided
- Sections B and C are coding questions that have to be answered in the system.

SECTION A – 20 MARKS

1	a)	What is Generative AI? Difference between discriminative and generative AI. (marks 2+ 3)	5
	b)	Explain the drawbacks of LSTM.	5
	c)	Draw the transformers architecture and explain the attention mechanism. (marks 3+4)	7
	d)	What is zero-shot learning?	3

SECTION B –40 MARKS

2		Use the data.csv dataset as provided in the notebook as pandas DataFrame and process it as questioned below.	
	a)	Pre-process the 'Text' feature as questioned below. • Remove the accented characters from the text feature. (3 marks) • Remove digits from the text feature. (3 marks) • Remove punctuations from the text feature. (3 marks) • Remove stopwords from the text feature. (3 marks) • Eliminate multiple spaces from the text feature. (3 marks) Note: Save this pre-processed text feature and use it as a feature for the next questions.	15
	b)	Find out the 5 most frequent words in the text corpus (from the preprocessed output of the previous question 2. a)	8

SRN □□□□□□□□□□□□□

	c)	Vectorize the pre-processed text feature by building/training a Skip-Gram Word2Vec model. Use this Skip-Gram Word2Vec model to fetch the top 5 most similar words for the word 'food'. (marks 3+5)	8
	d)	Vectorize the pre-processed text feature by building a CBOW Word2Vec model. Use the trained CBOW Word2Vec model to fetch the top 5 most similar words for the word 'food'. Is the output different from previous Skip-Gram's output? (marks 3+5+1)	9

SECTION C –40 MARKS

3	a)	Convert Textual output (of question 2. a) into numerical using countvectorizer	8
	b)	Convert Textual output (of question 2. a) into numerical using TfidfVectorizer	8
	c)	Build LSTM multiclass text classification model on the cleaned dataset (output of question 2. a) using Keras libraries.	20
	d)	show the confusion matrix and compute accuracy from the model and interpret it.	4