SRN | | | | | | | | | | | |

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

XXXX: 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
- Section B and C are coding questions which have to be answered in the system.

SECTION A – 20 MARKS

1	a)	What is Word Embedding? What is pre-Trained word embedding? What are the advantages of using pretrained word-embedding? (Marks-1+2+2)	5
	b)	What is Word2Vec. Explains different types/techniques of Word2Vec with example. (Marks- 1+4)	5
	c)	Define lemmatization and stemming. When stemming should be preferred over lemmatization? (Marks- 4 + 1)	5
	d)	Explain the RNN (Recurrent Neural Network) algorithm. What are the Key differences between RNN (Recurrent Neural Network) & LSTM (Long-Short term memory)? (Marks 3 + 2)	5

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)	Create a new panda DataFrame by fetching two columns 'text' and 'airline_sentiment' from data.csv. Use this new DataFrame for further processing as questioned below. Treat The 'text' column as feature column and 'airline_sentiment' as target column.	6

SRN | | | | | | | | | | | |

	b)	Clean the 'text' columns as questioned below. i. Convert all text to lower case. (Marks- 4) ii. Remove the URLs (http & www) from text. (Marks-6) iii. Remove stopwords from text. (Marks-8) iv. Remove punctuations from text. (Marks-8)	26
	c)	Fetch the top six most frequently used words from the text corpus.	8

SECTION C –40 MARKS

3		Use the cleaned DataFrame from previous section in order to build ML model as questioned below.	
	a)	Convert the cleaned text- column into numerical using Count-vectorization.	7
	b)	Convert the cleaned text -column into numerical using TF-IDF.	7
	c)	Split both Count-Vectorirzed and TF-ID dataset into train & test set with one fourth records being held for testing also ensure stratified sampling of target i.e., airline_sentiment on both splits. (3 +3)	6
	d)	Build a basic logistic regression model on Count-vectorize train set. Find out its accuracy on Count-vectorize test set. (5+5)	10
	e)	Build a basic logistic regression model on TF-IDF train set. find out its accuracy on TF-IDF test set. Which model has better accuracy? (5+5)	10