USN | | | | | | | | | | |

Internal Assessment Test 2 – November 2025

Sub	Natural Language Processing					Sub Code:	BCS714B	Branch :	CSE	
Date:	29/11/2025	Duration:	90 mins	Max Marks:	50	Sem /Sec:	VII Professional Elective		OBE	

	Answer any FIVE FULL Questions	MARKS	CO	RBT
1	Explain the working of the Naive Bayes classification algorithm with an example.	[10]	CO3	L2
2	Write short notes on 1. Training Naive Bayes classifiers 2. Optimizing for sentiment analysis	[10]	CO3	L2
3	Explain WordNet and FrameNet with its applications.	[10]	CO4	L2
4	Explain the cluster and fuzzy model of the IR system.	[10]	CO4	L2
5	What is machine translation? Explain word order topology in detail with an example.	[10]	CO5	L2
6	Describe a simple encoder-decoder model and identify the attention mechanism of the sentence "The girl liked the pink frock"	[10]	CO5	L3

CI CCI HOD-CSE

US N | | | | | | | | | | |

Internal Assessment Test 2 – November 2025

Sub	Natural Language Processing					Sub Code:	BCS714B	Branch :	CSE	
Date:	29/11/2025	Duration:	90 mins	Max Marks:	50	Sem /Sec:	VII Professional Elective		OBE	

	Answer any FIVE FULL Questions	MARKS	CO	RBT
1	Explain the working of the Naive Bayes classification algorithm with an example.	[10]	CO3	L2
2	Write short notes on 1. Training Naive Bayes classifiers 2. Optimizing for sentiment analysis	[10]	CO3	L2
3	Explain WordNet and FrameNet with its applications.	[10]	CO4	L2
4	Explain the cluster and fuzzy model of the IR system.	[10]	CO4	L2
5	What is machine translation? Explain word order topology in detail with an example.	[10]	CO5	L2
6	Describe a simple encoder-decoder model and identify the attention mechanism of the sentence "The girl liked the pink frock"	[10]	CO5	L3

CI CCI HOD-CSE

IAT-2

Q1:

A: Naive Bayes is a machine learning Algorithm that follows a probablistic model. Naive Bayes is a classification based model that is used for predictions. It is used to take decisions. Naive Bayes does not give a ranking but simply gives an yes or no answer.

This algorithm is based on the Bayes theorem and is given by the formula

$$P(c|x) = \frac{P(x|c) \times P(c)}{P(x)}$$

where $P(c|x)$ is the posterior ~~algorithm~~ probability,

$P(x|c)$ is the likelihood,

$P(c)$ is the prior probability,

$P(x)$ is the normalization constant.

According to Naive Bayes Algorithm there is an assumption that,

"All the features that belong to a class are independant of each other".

This assumption is made to solve the problem and predict the output.

The advantage of this output is that, the entire training process is made simpler and prediction of the output is much easier and faster comparing to the normal way.

The main disadvantage or challenge with the Naive bayes theorem is that, in most real world cases, the fact that all features ~~belong~~ are independant of each other is not true. Many features will be dependant or atleast correlated to each other.

According to the formula,

P(c) is the prior probability which is the frequency of the class.

P(x|c) is probability about how likely a particular feature will occur in that class.

There are mainly 3 steps to find the prediction:-

→ Find the prior probability P(c) and P(x|c) which is the likelihood.

→ Find the posterior probability P(c|x).

→ The class with the highest posterior probability is the answer / output.

For eg.

Let us use Naive Bayes to find out whether an email is spam or ham.

→ email1, email2 : spam
→ email2, email2 : ham
→ email1, email1 : spam

Prediction: email2, email2

→ find prior probability

P(spam), P(ham)

↓ 2/3 ↓ 1/3

→ Find likelihood,

$P(email2/spam) = 1/4$ $p(email2/ham) = 2/2$

$P(email1/spam) = 3/4$ $p(email1/ham) = 0$.

→ If posterior of spam is greater than ~~P~~(spam) is the output or ham is. $P(email2/spam) = \frac{2}{3} \times \frac{1}{4} = \frac{1}{6}$.

Q2. $p(email2/\cancel{spam}\ ham) = \frac{2}{2} \times \frac{1}{3} = \frac{2}{6} = \frac{1}{3}$. since, $\frac{1}{3} > \frac{1}{6}$,

ans $\boxed{\frac{1}{3}}$ $\boxed{\text{ham email}}$

1A: Training Naive Bayes Classifier.

→ Naive Bayes is a probablistic model that is used in classification to predict the output.

The following steps are used in training the classifier:-

→ Collecting Data:-

The first step is too collect the data that is relevant to the prediction. The collected data should be made sure that it is labeled in accordance to the classes.

eg. "The movie is good" — positive

"The movie is bad" — negative

- → preprocessing data :-

The data that is collected has to be preprocessed for information and has to be cleaned.

→ The data has to stripped used stemmers, to make it a root word.

→ All stop words and punctuations have to be removed.

- → calculating prior probability :-

From the data recieved based on the frequency of classes the prior probability of each class has to be found.

- → calculating likelihood :-

From the data based on the frequency of features that belong to each class the likelihood of the feature in the class has to be discovered.

- → Handling Imbalance in class :-

If one particular feature is present in the test data but not in the training data, that feature can be dealt with using laplace smoothing also known as add on smoothening.

- → Calculating posterior data:-

Using the formula, posterior data is calculated and the classification is made.

2. Optimizing for sentiment Analysis can be done using the following steps:-

- → Feature Engineering:-

The sentence taken rather than being worked on as a whole can be worked on by splitting it into parts using

- → unigram: one word splits
- → bigram: 2 word splits
- → n-gram: n word splits.

when a sentence is split it is easier to understand by a system rather than a whole.

- → Data Preprocessing:-

The data has to be normalized and cleaned to handle ~~[illegible]~~ the training. the following steps will optimize the analysis:-

- Remove all stopwords (eg. etc, i.e)
- Remove all punctuations.
- convert the sentence to lowercase.

- → Handling Imbalances:-

The Data recieved will always carry imbalances like one class will have more data whilst the other might have less data. It is important to balance this data for better outcomes.

- → Cross validation and hyperparameter evaluation:-

The data is then divided into multiple parts to check for accuracy and

the parameters are tuned in a way to achieve better results.

• → Model evaluation :-

The model must be checked for accuracy, precision, recall, AUC, F1-score and various other parameter to check legibility of the model.

10

Q3.

A: WordNet is a lexical Database that is used worldwide for Natural language processing and Information Retrieval. It was developed by the "Princeton University" in the United States.

According to this database, similar words are put in groups called "synsets" and their are links established between these words using relationships.

these relationships include :-

- hyponomy (is-a relationship)
- hypernomy (has-a relationship)
- meronomy (part-whole relationship).

As these relationships are built across these words it becomes easier to draw lines between them and use them.

For eg. "Dog" and "canine".

these two words are synets.

"Dog" and "Animal".

Dog has a hyponomy relationship with Animal.

It is an animal.

The applications of word Net include :-

→ wordsense Disambiguity :

A word can be disambigued easily used wordnet. For eg. the difference between a "bank" and a river "bank" can be drawn

using relationship links.

- → Machine Translation :-

Translating one language to english, or vice-versa can be done with the help of word net.

- → Text classification :-

Text can be classified into different categories using wordnet.

- → ~~sp~~ Sentiment Analysis :-

Word net is widely used for sentiment analysis ~~and~~ as it can predict relationship between words.

FRAME NET

→ Framenet is a lexical database that used word wide for its information. It was developed in Berkely and has been used in research of NLP and Information Retrieval.

In Framenet, a sentence is divided into frames, each frame ~~has~~ is typically one word.

Each frame has its own participants, props and responsibilities.

Using these frames, links are made to other words to depict relationships.

For eg. "A robber was arrested"

Let "arrest" be the frame:

the various participants will be -

- policemen
- Robber.

props will be:-

- handcuffs
- jail

Responsibilities:-

- catch the criminal.

20

These frames make it easy to draw links.
The various applications of Framenet are:-

- → Information Retrieval:-

With the help of these frames, retrieving information has become easier.

- → Speech Recognition:-

It is easier to recognize speech or convert speech to text based on these databases.

- → Word Disambiguity:-

It is easier to understand the meaning of words that are same in spelling.

- → Text Classification:-

Classify text to the class they belong has also become easier.

- → Machine Translation:- With large data that has become widely available it is easier for machines to train and translate text.

Q4

A: Cluster model

The Cluster model of an Information Retrieval System is a type of model based on vector space. In this model groups of documents are made based on similarity, so that document finding becomes easier.

A cluster of similar documents are made. Each document is converted to a context vector in a multidimensional space where the dimensions corresponds to a word / term.

The similarities between documents are found using mathematical algorithms such as cosine similarity or Euclidean Distance. The clusters are the made using K means algorithm or hierarchial algorithm.

This is an efficient ways of retrieving similar documents rather than just retrieving one.

The various advantages of Cluster Model are :-

- Improved Search Results : Each search results in only similar documents therefore improving performance.

- Easy Retrieval :- Retrieving data has become easier with these models.

Disadvantages :-

- Cluster Quality : The Quality of the cluster majorly depends on the algorithm used.

- Dynamic dataset : With changing datasets it becomes difficult to form correct clusters.

Fuzzy Model

The Fuzzy model uses Fuzzy logic ~~of the~~ for information Retrieval. According to the Fuzzy logic, rather than just deciding

whether a document is relevant or no it is important to give a degree of relevance to the documents.

The fuzzy model unlike boolean does not just answer yes or no (0 or 1) but gives a range between 0-1 to depict relevance.

For eg. A document has a relevance degree of 0.7, it might not be the but document / information for the problem but gives ~~a range~~ decent information to the user.

The advantages of the fuzzy model are:-

- → Flexibility:-

There is no yes/no case here, the fuzzy model increases flexibility for the user to choose from.

- → Relevance:-

The fuzzy model gives a bunch of documents that are relevant to the user

to choose from.

The disadvantages of the Fuzzy model are:-

- → Complexity:

The model is more complex than a base model like boolean for information retrieval.

- → Interpretability:

Many users will not understand the relevance degree to make a decision out of it.

10

Q5.

A:- Machine Translation is the concept of using a computer software to translate one language to another. Humans are capable of doing translation but it is very important in this day and age, for a machine to help humans understand languages that they do not know. There are various models ~~are~~ that are used for machine translation

• → Rule based ~~model~~ machine translation :-

— This rule based model knows all the rules of both the languages for translation, and translates based on these rules.

— These rules are researched and written by human experts in the field.

• → Statical based machine translation :-

— This model assumes there is a statistical underlying probabilitic distribution.

— This model translates based on this distribution and looks at the words probability being a part of the sentence.

• → Neural Networks based machine translation :-

— This model using deep learning techniques like RNN (Recurrent Neural Network) and LSTM (Long short term memory) to

learn the ~~na~~ dataset and is able to efficiently translate languages.

Word Order Topology

Word Order Topology is a concept where ~~wot~~ the sequence of the words in a sentence is decided. The sequence of the 'subject, object and verb' depends on the language it is written in. the translation between these languages need to be mindful of this order.

The various sequences are:-

- Subject Verb Object [SOVO] :

  This order is typically followed in the engish language and is the most easiest

  For eg :

  $\underbrace{\text{The girl}}_{\text{subject}}$ $\underbrace{\text{like}}_{\text{verb}}$ $\underbrace{\text{apple}}_{\text{Object}}$

- Subject object verb :-

This order is followed in Asian languages like Japanese, korean where the subject object and the verb comes.

eg.

the girl	apple	likes
subject	object	verb.

- verb subject object :-

This sequence is followed in morphologically rich languages like Arabic and many Indian languages where char based translation take place.

eg.

likes	the girl	apple
verb	subject	object.

Neural Networks can easily predict the word order topology of sentences during translation.