Mid Semester Examination
Course Name: Natural Language Processing
Code: CS 563
Full Marks-40 **Time: 2 hours**
Answer ALL the questions

Make reasonable assumptions as and whenever necessary. You can answer the questions in any sequence. However, answers of all the parts to any particular question should appear together.

(Q1). (a). Formulate Named Entity Recognition (NER) using Maximum Entropy Markov Model. Tag the following text manually with named entity (NE) classes: PER (*Person)*, LOC (*Location)*, ORG (*Organization*), DT (*Date)*, TE (*Time*), NUM (*Number*) and O (*Others)*.

*"**As reported by TOI on Saturday,** Rohit Sharma **will lead India's 15-member squad in the** Nidahas Trophy **T20 tri-series starting March 6 in Sri Lanka. Top stars like** Virat Kohli, MS Dhoni, Bhuvneshwar Kumar, Jasprit Bumrah **and** Hardik Pandya **have been given rest for the event while** Shikhar Dhawan **has been named the vice-captain. Bangladesh is the third team in the event".***

Use BIO (Beginning, Intermediate and Outside) notation for NE tagging. What can be the potential features for NER? Define the possible non-admissible sequences, and formulate a probabilistic framework to reduce these errors (*with proper justifications*).

(b). Explain with proper examples why is local contextual information not sufficient for detecting NE properly? How can global contextual information be incorporated into the classification model? **3+5+5+3+5**

(Q2). Consider the following example:

Maradona's vision, passing, ball control, dribbling skills, speed, reflexes and reaction time was combined with his small size (1.65 m or 5 ft 5 in tall) giving him a low center of gravity which allowed him to maneuver better than most other football players; he would often dribble past multiple opposing players on a run.

Assume you have only 5 tags : *N (noun), V (verb), J (adjective), R (adverb) and F (other, i.e., function words).* Manually Part-of-Speech (PoS) tag the above text. Use the convention ***'word_POS'***. What possible sequences of tags will be asssigned to the above text ? Answer this question with respect to a second order Hidden Markov Model (HMM) based sequence learning algorithm. Use the concepts of State Sequence Probability (decoding using Viterbi), Observation Probability (forward and bacward probabilities), and HMM model parameters etc. **3+5**

(Q3). (a). How do feature selection and ensemble learning improve the efficiency of classification? "Determining weights of classifiers in an ensemble framework can be modeled using Genetic Algorithm"- Formulate this problem, and show the necessary steps with proper explanations.
(b). Consider the examples in **Question 1** and **Question 2** as two documents. Design a Vector Space Model using these two documents by eliminating stop words, performing stemming, creating term-document matrix with tf-idf weighting scheme. **2+3+6**