Q 1.7

1.
The question glosses over the selection of the corpus. The choice of text corpus for analysis greatly influences the entropy value. For instance, a choice of Brown corpus, a balanced corpus containing a wide range of text types, will yield different entropy values compared to a corpus of modern-day English tweets that is informal and conversational.

The question doesn't provide an utterance overlooking the impact of context. The probability of a word occurring is related to words that precede and follow it, which indicates that per-word entropy is not constant across all utterances. Ignoring the context dependence can lead to an oversimplified and potentially misleading calculation of per-word entropy.

The question overlooks the selection of language model. Using different language models can generate different entropy values. Specifically, the entropy calculated using a unigram model would be different from that calculated using a bigram model, as the latter model incorporates more information about the structure and patterns of language.

2.
Firstly, I will choose a large and diverse corpus, such as the Brown corpus, as the training data. The diversity of the corpus ensures that various forms of language (formal, informal, technical, etc.) are included.

Secondly, I will select a bigram language model for the experiment. Compared to unigram, bigram strikes a balance between complexity and capturing sufficient context. Specifically, bigram provides more accurate word probabilities than unigram as it considers the preceding word, making it better suited for estimating entropy in a naturally contextual language like English.

Thirdly, process the data to train the model. I will convert the text in the corpus into tokens (words). This process includes handling punctuation, capitalization, and potentially filtering out non-words (like numbers or symbols).

Afterwards, calculate the word frequencies and probabilities. I will count the frequency of each word and each bigram in the corpus and convert them into probabilities. For a bigram model, the probability of a word $$ \ w \ $$ given its preceding word $$ \ w_{prev} $$is calculated as $$ \ P(w|w_{prev}) = \frac{\text{Count}(w_{prev}, w)}{\text{Count}(w_{prev})} \ $$. For each bigram $$ \ (w_{prev}, w) \ $$, calculate the joint probability $$ \ P(w_{prev}, w) \ $$. $$ \ P(w_{prev}, w) = \frac{\text{Count}(w_{prev}, w)}{\text{Total Bigrams}} \ $$

Finally, use the formula: $$ \ H = -\sum_{w_{prev},w} P(w_{prev}, w) \log_2 P(w|w_{prev}) \ $$ to calculate entropy in each bigram, then sum these values across all bigrams. Divide the total entropy by the number of bigrams to get the average bigram entropy. Then, divide this average by 2 to estimate the per-word entropy of the Brown corpus (English).

3.
The evolution of language is often driven by cultural and societal changes. Words can gain new meanings, lose old ones, or shift in connotation based on societal trends, technological

advancements, and other factors. Additionally, as words evolve, their frequency in the language can fluctuate. A word might become more common as it gains a new meaning or less so if its original context becomes outdated.

To capture these changes, I will choose a corpus that spans different time periods, ideally one that is designed to reflect language changes over time. It allows us to observe shifts in word usage and frequency across different eras.

Q 2.2

1.
The individual features provide basic information about the sentence structure but can not capture the relationships between these phrases. When combining these features ([V=N1=P=N2]), there's a significant increase in accuracy, which suggests that the interaction between the phrases is critical for predicting the correct class. It implies that the task requires how these elements relate to each other within the sentence structure to make accurate predictions.

2.
The Naive Bayes model accuracy is 79.49987620698192%, slightly lower than the logistic regression. Naive Bayes assumes all the features are independent given the class label. Logistic regression captures the iteration between features.

3.
I would be against it. A binary feature captures a rule that may strongly indicate the correct classification. It may not apply to sparse data if the rule is too specific. Additionally, if these conditions rarely occur together, the feature may not contribute much to the model's performance.

Q 2.3

1.
Basic features are the individual headwords (‘verb’, ‘np1’, ‘prep’, ‘np2’). Pairwise combinations (‘verb_np1’, ‘verb_prep’, ‘verb_np2’, ‘np1_prep’, ‘prep_np2’, ‘np1_np2’) capture binary relationships between the headwords. Three-word Combinations combine three headwords at a time (‘verb_np1_prep’, ‘np1_prep_np2’), which provide context that may be crucial for understanding more complex syntactic structures that influence PP attachment. Common Verb-Preposition Combinations(‘common_verb_prep_combo’) look at specific verb-preposition pairs that are known to have strong attachment preferences based on syntactic conventions or collocation frequencies.

2.
‘prep:of==1 and label is 'V’’: A negative weight for this feature in the context of a 'V' label implies that when "of" is present, it is a strong predictor for noun attachment. It means the model has learned from the training data that PPs starting with "of" are usually modifying a nearby noun instead of the verb.

The feature ‘verb_prep:giving_to==1 and label is 'N'’ indicates that when the verb "giving" is followed by the preposition "to," this combination is a strong predictor for the prepositional phrase (PP) attaching to a noun (‘N’). In English, the construction "giving to" often introduces a PP that is more closely associated with a noun that follows it. It is useful for a classifier focused on disambiguating PP attachment because it captures a specific and common verb-preposition combination that has a clear tendency in terms of syntactic attachment.

The feature ‘verb_prep:are_of==1 and label is ‘V’’ with a significant weight suggests that when the verb "are" is followed by the preposition "of," this combination is a strong predictor for the prepositional phrase (PP) attaching to a verb (`V`). "are of" commonly appears in verb phrases that are followed by a PP providing further detail or clarification about the verb or the overall action. It implies that in contexts where "are of" appears, the classifier has learned that the preposition is more likely to be part of a phrase that modifies or extends the verb rather than a noun.