Foundations of Intelligent Systems **Name : Pillai Rajkumar Lenin**
LAB-2 Report **Section-01**
Prof : Jansen Orfan **Course: CSCI-630**

1. **Following distinct features were considered for the decision tree.**

Features	Meaning
• Contains-de	The Dutch translation of English word the is de ,this feature checks whether any word de is present in the sentence. As the is most commonly word is used in a sentence so this was considered as one of the features.
• No of ij	This feature checks whether or not any sentence with substring ij is present, since ij is most frequently used word in Dutch and in English not many words have ij as substring.
• No of es	This attribute the no of e's present in the given sentence. If it is more than 13 then it classifies it as Dutch. It has been found that the no of e used in Dutch sentence is between 13-19 and in English sentence is 8-12.
• Containshet	This attribute is used to check if the word "het" is present in sentence and classifies it as a Dutch sentence. "het" also refers to English word "the " in some sentences
• Containsvan	This attribute checks if the word "van" is present and classifies it as Dutch. The word "van" refers to from in English language, so it is also frequently used word in a sentence.
• Containsoo	This feature checks whether or not any sentence with substring oo is present, since oo is found frequently in the Dutch sentences.
• Containsee	This feature checks whether ee is present as a substring in sentence and classifies it as Dutch language as it English senetences less likely have substring ee compared to dutch.

Every attribute above is a Boolean attribute, where True refers to a Dutch sentence and False refers to English sentence.

2. Decision Tree Learning:

The decision tree implemented in this program is based on the Entropy and Information Gain algorithm. The entropy for a attribute is calculated using the following formula in the program:

result= ((((Pvalueone+Nvalueone))/((encounter+nlcounter)))*(IPN(Pvalueone,Nvalueone))) + ((((Pvaluetwo+Nvaluetwo))/((encounter+nlcounter)))*(IPN(Pvaluetwo,Nvaluetwo)))

In the above equation The IPN function is defined as follows:-

answer= ((-Pvalue/(Pvalue+Nvalue)))* math.log2((Pvalue/(Pvalue+Nvalue))) - ((Nvalue/(Pvalue+Nvalue)))* math.log2((Nvalue/(Pvalue+Nvalue)))

Where Pvalue refers to no of example with True value for one label of target attribute and Nvalue refers to no of example with False value for other label of target attribute

The gain is calculated by following formula:
Gain=(entropyforclass- entropyforattribute)

In this program a gaindictionary is used to store the values of gain for each attribute during each iteration and the attribute with maximum Information gain is considered for splitting.

The following attribute had the maximum Gain among all other attributes and so it is the first attribute considered for splitting ,hence it forms the root node of decision tree.

'containsde': 0.4678234948275535

In further iterations the attributes except the one which has already been selected for split were considered for split and the decision tree obtained is as follows:

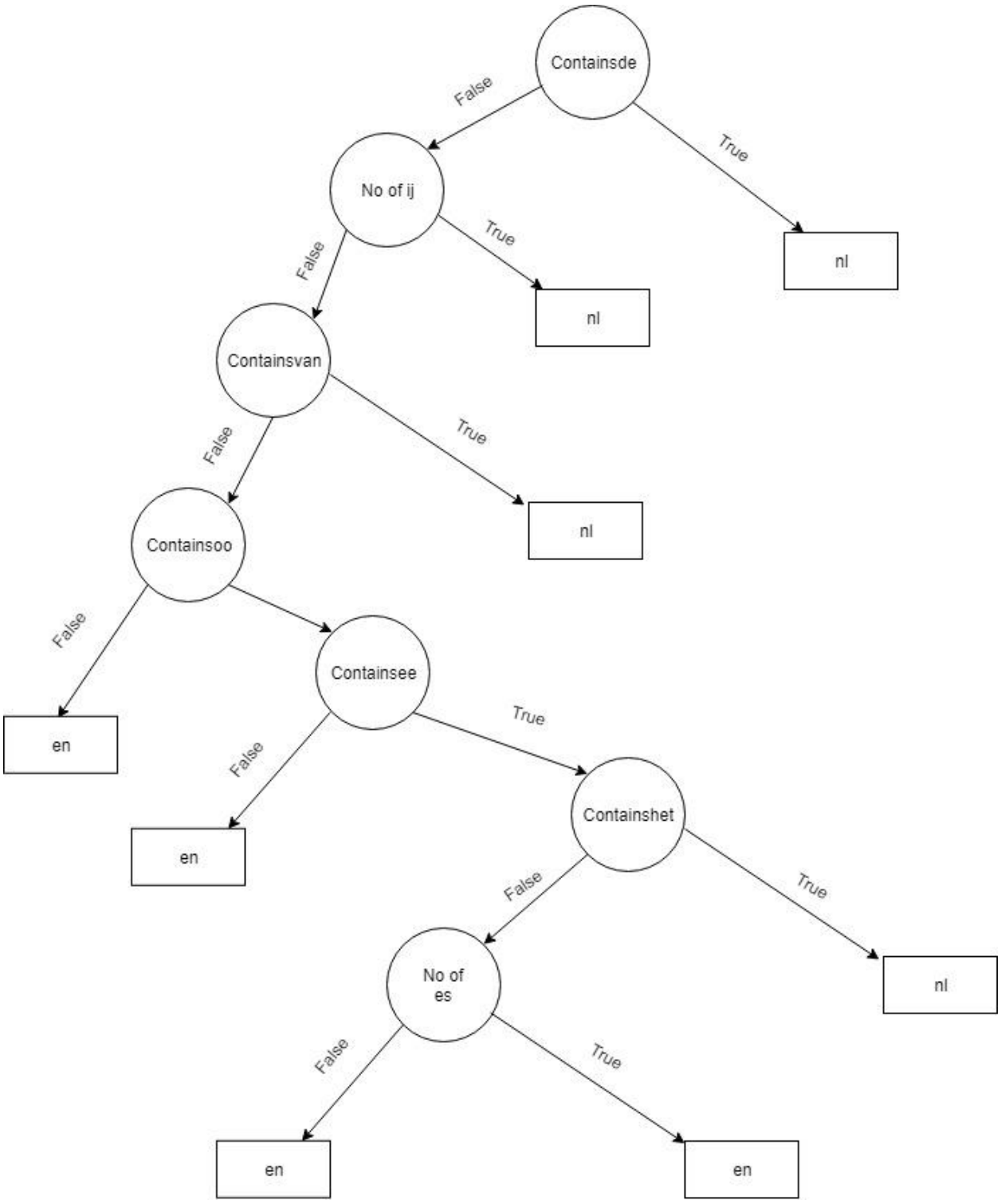

Fig - The decision tree built using the training set

The training set is loaded from files english.txt and dutch.txt which contains the sentences in English and Dutch language respectively. The sentences were collected from English and Dutch Wikipedia and also from various English and Dutch Essay websites.

The decision tree is stored in a dictionary and is written to a file name specified while running the program

The file train.py is the program which performs the training and decision tree can be seen as a output in a separate file.

The prediction is performed by file predict.py which contains the decision tree model built on the basis of training data and predicts the language of the sentences based on this decision tree. The output is stored in a separate file with the appropriate English and Dutch labels.

The training and the testing of the data can be done in following manner:

```
rlp4867@glados:~/MyCourses/FIS$ python3 train.py
Please enter filename with dutch samples:'dutch.txt' dutch.txt
Please enter filename with english samples:'english.txt' english.txt
Please enter filename to store the decision tree:'hypothesisout.txt' hypothesisout.txt
The decision tree built is stored in:  hypothesisout.txt file
rlp4867@glados:~/MyCourses/FIS$ python3 predict.py
Please enter filename which conains the sentences:'test.txt' :test.txt
Please enter filename in which the predictions are to be stored :output.txt
nl
en
en
nl
en
en
nl
en
nl
en
The output of predictions is stored in:  output.txt file
```