Complex word classification

1.Modelul K-Nearest Neighbors (KNN)

Pentru modelul KNN am folosit caracteristici cu privire la structura cuvantului, la cuvant si la corpus.

Caracteristicile folosite sunt:

- O functie care verifica daca cuvantul apartine listei de cuvinte Dale Chall si returneaza 1 daca apartine respectiv 0 daca nu apartine

-O functie care numara aparitiile fiecarui cuvant in textele de antrenare si de testare si returneaza numarul acestora \* 100 / numarul de cuvinte unice

-O functie care returneaza numarul de litere mari din cuvant / 10

- O functie care verifica daca vucantul contine litere rar intalnite (‘jqzx’) si returneaza 1 daca contine / 0 daca nu contine.

- O functie care numara sensurile din dictionar ale cuvantului si returneaza numarul acestora / 10

- O functie care diferentiaza cuvintele prin corpusul din care fac parte

Hiperparametrul folosit pentru model este k=1. Pentru a alege cel mai bun hiperparametru am imartit datele in date de antrenare si date de validare. Am testat hiperparametri k = 1-20 si am calculat acuratetea balansata pe datele de validare. Am ales hiperparametrul cu care am obtinut cea mai buna acuratete.

Antrenarea dureaza cateva secunde.

Performanta obtinuta pe datele publice este de : 0.59350.

2.Modelul Naive-Bayes

Pentru modelul Naive-Bayes am folosit caracteristici cu privire la structura cuvantului, la cuvant si la corpus.

Caracteristicile folosite sunt:

- O functie care verifica daca cuvantul apartine listei de cuvinte Dale Chall si returneaza 1 daca apartine respectiv 0 daca nu apartine

- O functie care verifica daca cuvantul este titlu si returneaza 1 daca este, iar in caz contrar 0

- O functie care numara literele mari din cuvant si returneaza numarul acestora / 10

- O functie care numara vocalele din cuvant si returneaza numarul acestora / 10

- O functie care numara cifrele din cuvant si returneaza numarul acestora

- O functie care verifica daca vucantul contine litere rar intalnite (‘jqzx’) si returneaza 1 daca contine / 0 daca nu contine.

- O functie care numara sensurile din dictionar ale cuvantului si returneaza numarul acestora / 10

- O functie care diferentiaza cuvintele prin corpusul din care fac parte

Modelul nu a folosit hiperparametrii.

Antrenarea dureaza cateva secunde.

Performanta obtinuta pe datele publice este de : 0.75268.