Componenta C

Set de date: <https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned>

* 7650 entries
* Label-uite cu 0 sau 1 pentru nu e depresie sau e
* Preprocesare: LIWC tool categorizeaza cuvintele din input

Metrici de evaluare:

* Recall,Precision,Accuracy,F1-score
* Confusion Matrix
* Area under Receiving Operating characteristic curve(graph showing the performance model at all classification threshold. Two parameters: True positive rate, false positive rate)

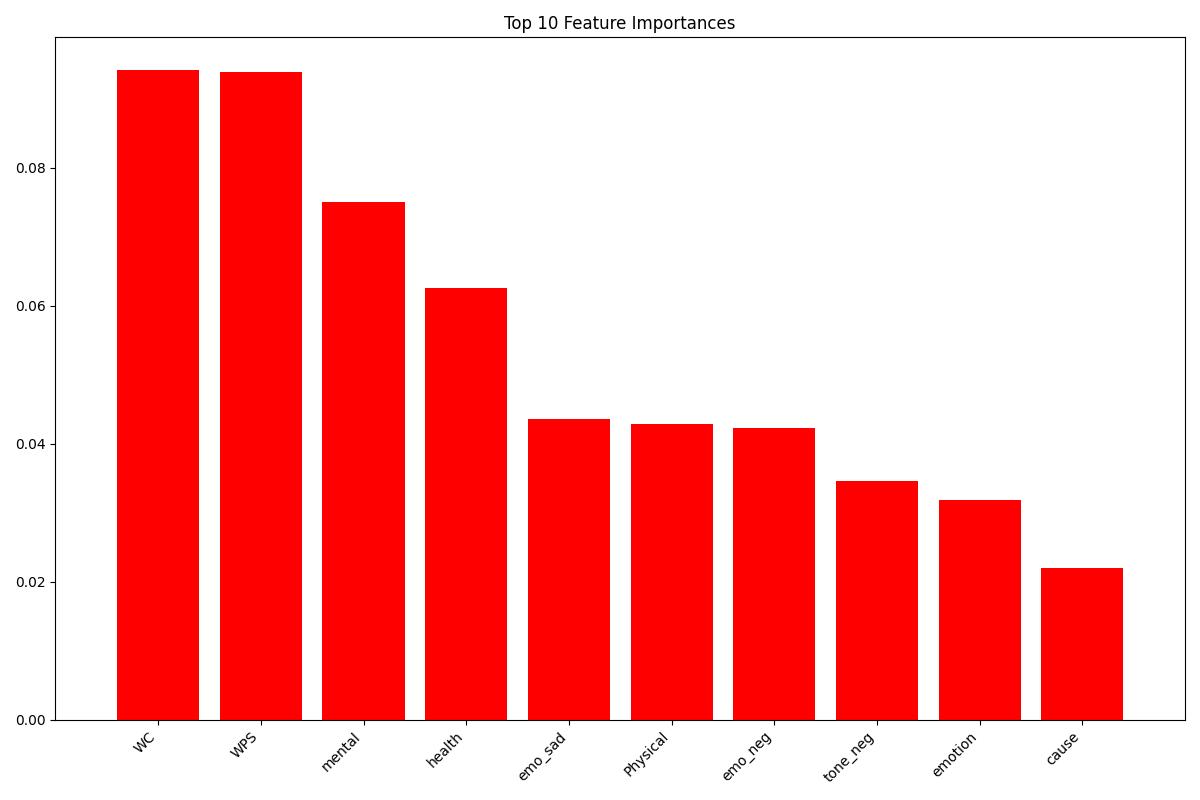
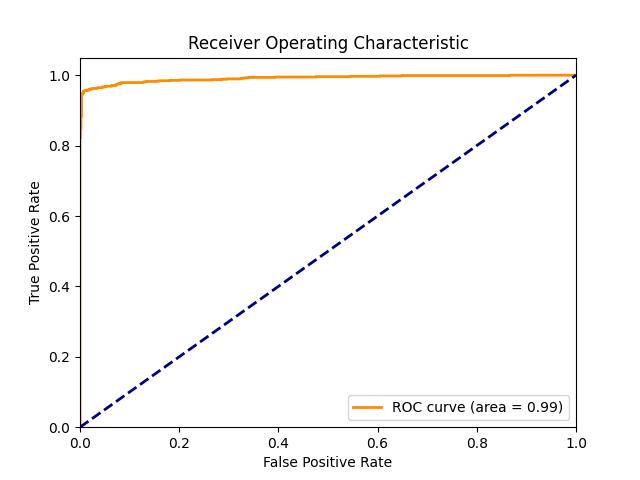
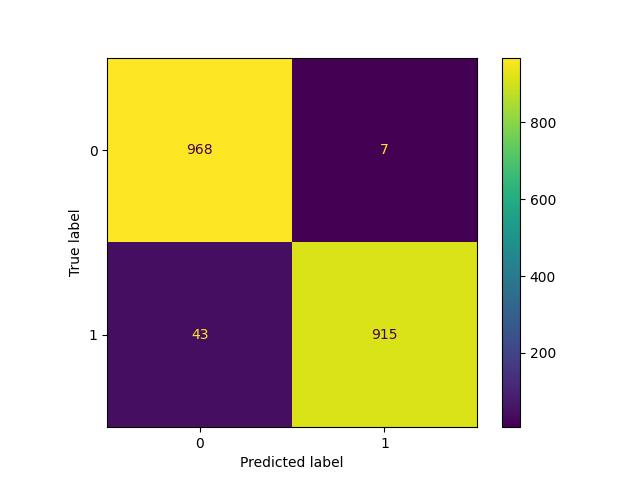
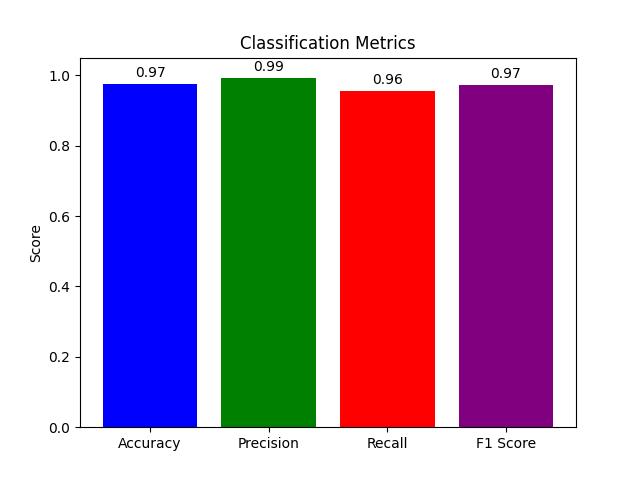
Rezultate:

1. Model in engleza:

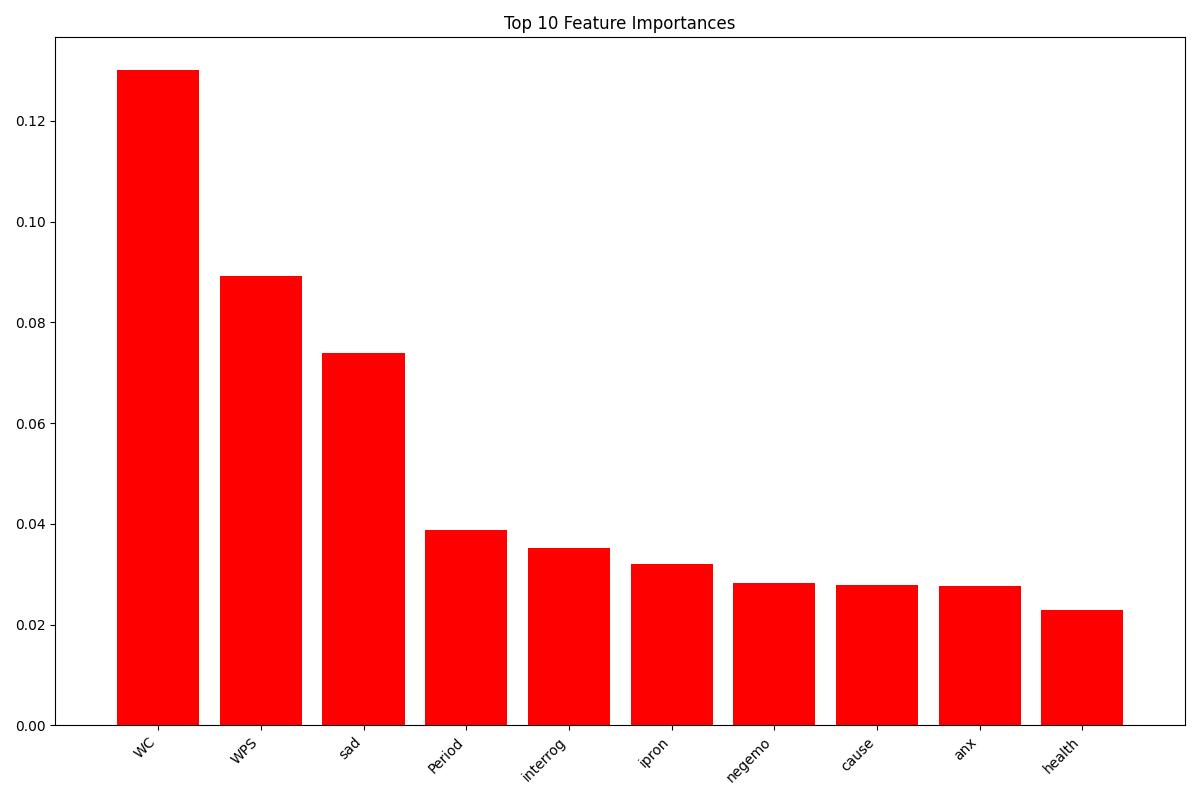
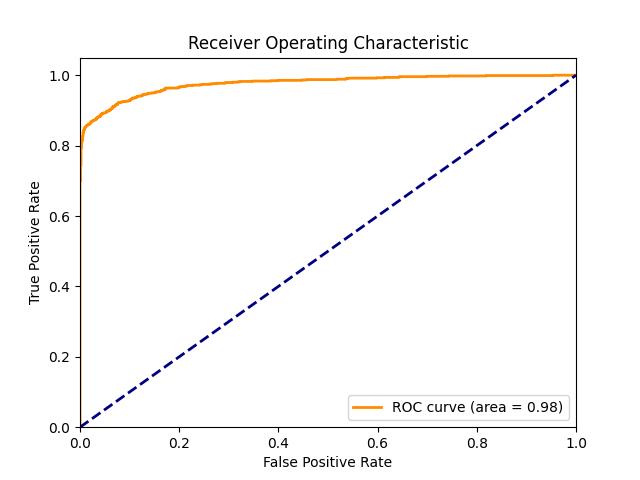
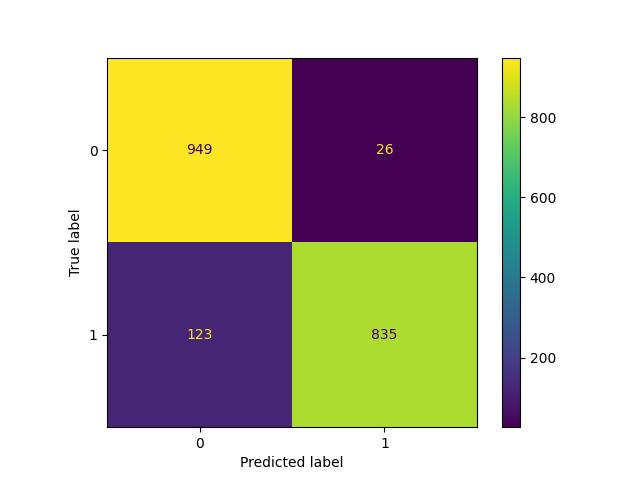
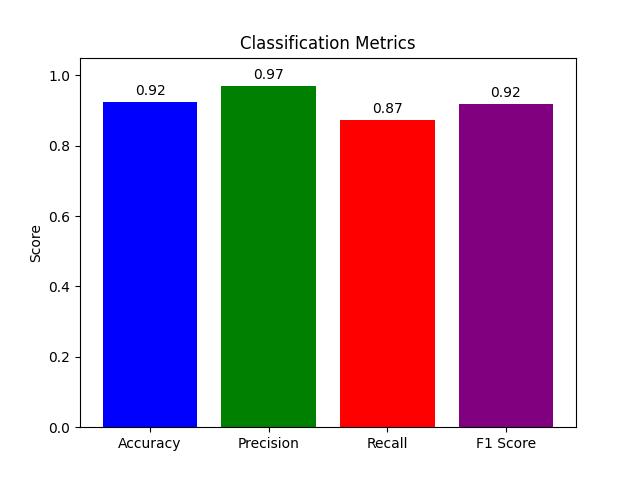
Analiza:

Probleme apar la detectarea positivelor.

Word Count si Words per sentence sunt cele mai importante feature-uri.



1. Model in Romana



Analiza:

Metricile de clasificare sunt mai slabe decat cele de la engleza

Acceasi problema de identificarea positivelor este si mai evidenta

Sunt si mai importante WC si WPS, ceea ce duce la o performanta mai slaba

Detalii antrenare:

* 75/25 train/test split
* **mtry**: Acesta reprezintă numărul de caracteristici luate în considerare pentru divizarea la fiecare nod. Valorile mai mici promovează diversitatea arborilor și sunt benefice atunci când există mulți predictori relevanți. Având în vedere setul mare de caracteristici, se sugerează o valoare mai mică decât rădăcina pătrată a numărului de caracteristici pentru a preveni ca anumite caracteristici dominante să umbrească pe celelalte.
* **Numărul de arbori**: Un număr suficient de arbori asigură predicții stabile și estimări corecte ale importanței. Deși mai mulți arbori îmbunătățesc în general performanța modelului, dincolo de un anumit punct câștigurile marginale scad. În scopuri practice, se recomandă între 500 și 1000 de arbori.
* **Dimensiunea nodului**: Dimensiunea nodului controlează adâncimea arborelui. Dimensiunile mai mici ale nodului pot duce potențial la supraînvățare, în special atunci când numărul de caracteristici este mare. O dimensiune a nodului mai mare de 1 este preferată pentru a atenua acest risc și pentru a îmbunătăți eficiența computațională.
* **Dimensiunea eșantionului**: Proporția de date utilizată pentru antrenarea fiecărui arbore. Dimensiunile mai mici ale eșantionului duc la mai multă diversitate, dar pot scădea acuratețea fiecărui arbore individual. Dimensiunea optimă a eșantionului trebuie să fie specifică problemei, dar eșantionarea unui subset, cum ar fi între 20% și 90% din date, poate produce rezultate bune reducând în același timp timpul de rulare.
* **Valorile alese**: 6 pentry mtry, 900 pentru number of trees, 3 pentru node size and 8 pentru ample size.