Introducere

În cadrul temei 3, am implementat un clasificator și sumarizator de articole, având la bază algoritmul *Naive Bayes*.

Pentru acesta, am folosit articolele puse la dispoziție în arhiva temei, încărcându-le în memorie sub forma unei liste, denumită *articles*.

Detalierea structurii acestuia:

articles – listă cuprinzând tot setul de documente, având câte 5 subseturi, de tip dicționar, reprezentând categoriile de știri, a cărui implementare este descrisă mai jos:

articles = [

{

”category”: ”business” / ”entertainment” / ”politics” / ” sport” / ”tech”

”articles” : [listă de elemente tip article]

}

]

article – dicționar în care sunt păstrate elementele necesare algoritmului: textul inițial, id-ul fișierului, numele fișierului, sumarul fișierului (cheile : ”text”, ”id”, ”name”, ”summary”), iar mai apoi elementele de procesare pe text:

- ”tokens” : tokenizarea textului inițial al articolul
- ”no_stopwords” : îndepărtarea cuvintelor de tip *stopword* din tokenuri
- ”lematized” : lematizarea tokenurilor rămase după îndepărtarea stopwords.

Pași similari au fost realizați și pentru partea de sumarizare, aplicându-se aceleași procedee și pentru sumarul atribuit textului.

Clasificare

Pentru a testa capacitatea de clasificare a modelului, am rulat de mai multe ori, pentru fiecare dintre cele trei variante în parte:

1. Varianta tokenized, cu stop words incluse:

   Valorile obținute au fost:

   Precision: 0.9368

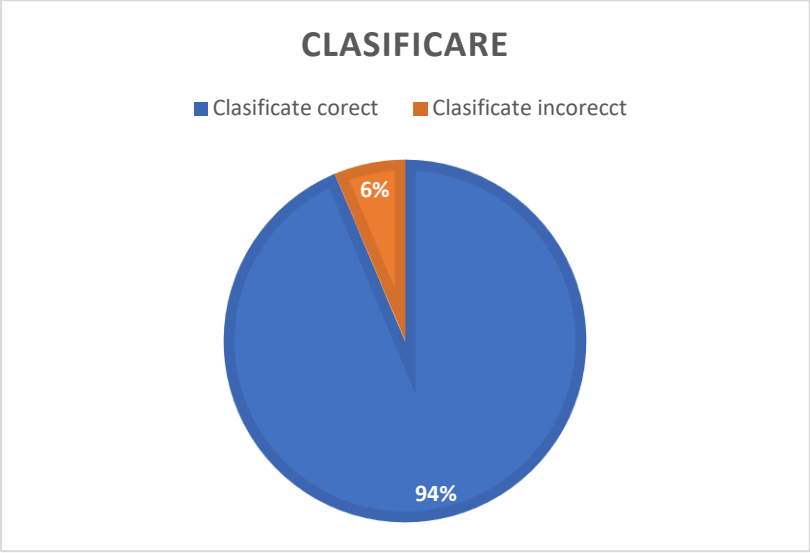
   Recall :

   - Business : 0.9606
   - Entertainment : 0.6875
   - Politics : 1.0
   - Sport : 1.0
   - Tech : 1.0

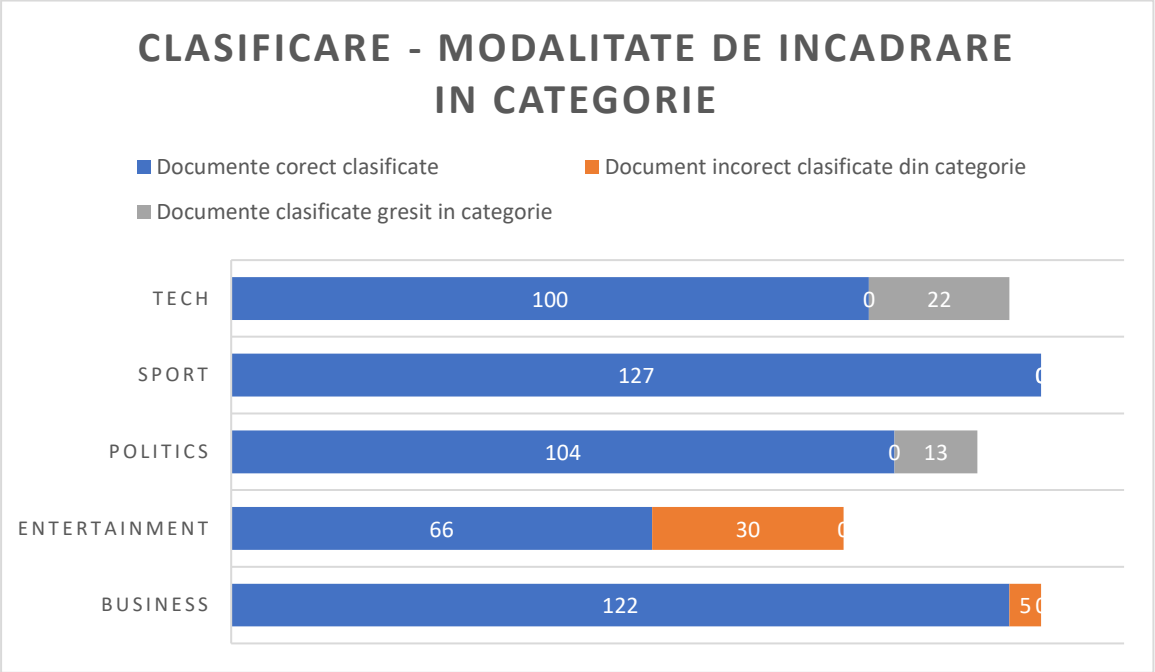
   Matricea de confuzie corespunzătoare:

	Business	Entertainment	Politics	Sport	Tech
Business	**122**	0	3	0	2
Entertainment	0	**66**	10	0	20
Politics	0	0	**104**	0	0
Sport	0	0	0	**127**	0
Tech	0	0	0	0	**100**

Procentajul de articole clasificate corect este de 94%, un procentaj ce poate fi considerat bun pentru algoritmul utilizat.



Clasificarea, pe categorii de articole, indică faptul că niciun articol de tip „tech”, „sport” sau „politics” nu a fost încadrat greșit în altă categorie. De asemenea, majoritatea articolelor care au fost clasificate greșit în această variantă au ajuns să fie clasificate ca fiind „tech”, următoarea opțiune fiind „politics”. In categoria „sport”, au fost identificate toate și numai articolele aparținând acestei categorii.



2. Varianta tokenized, fără stop words

   Valorile obținute au fost:

   Precision: 0.9404

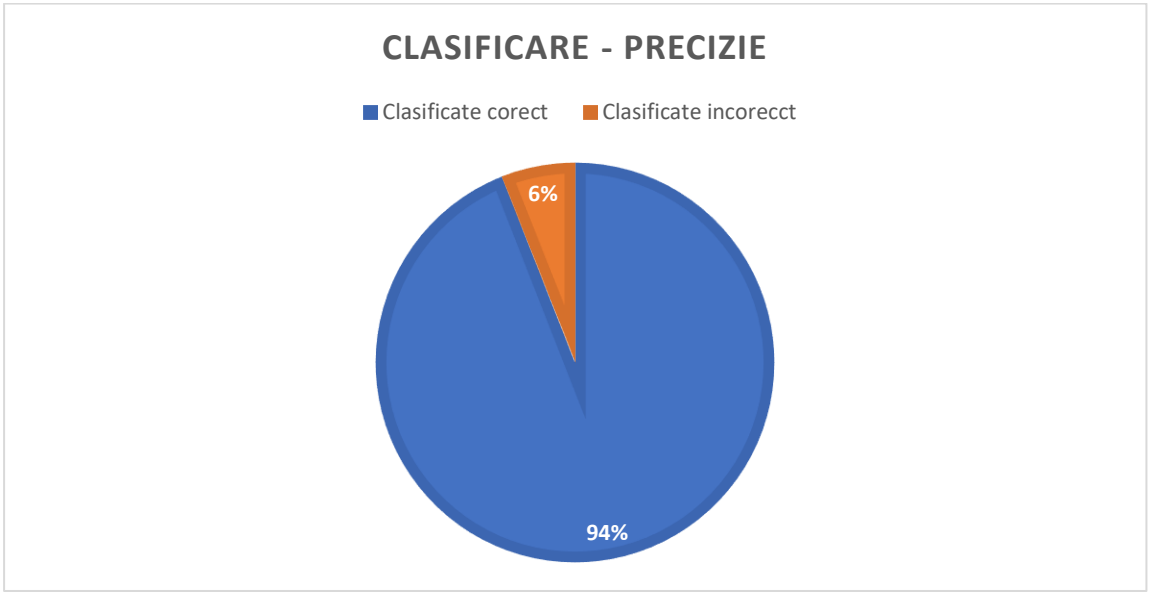
   Recall :

   - Business : 0.9370
   - Entertainment : 0.7708
   - Politics : 0.9807
   - Sport : 1.0
   - Tech : 1.0

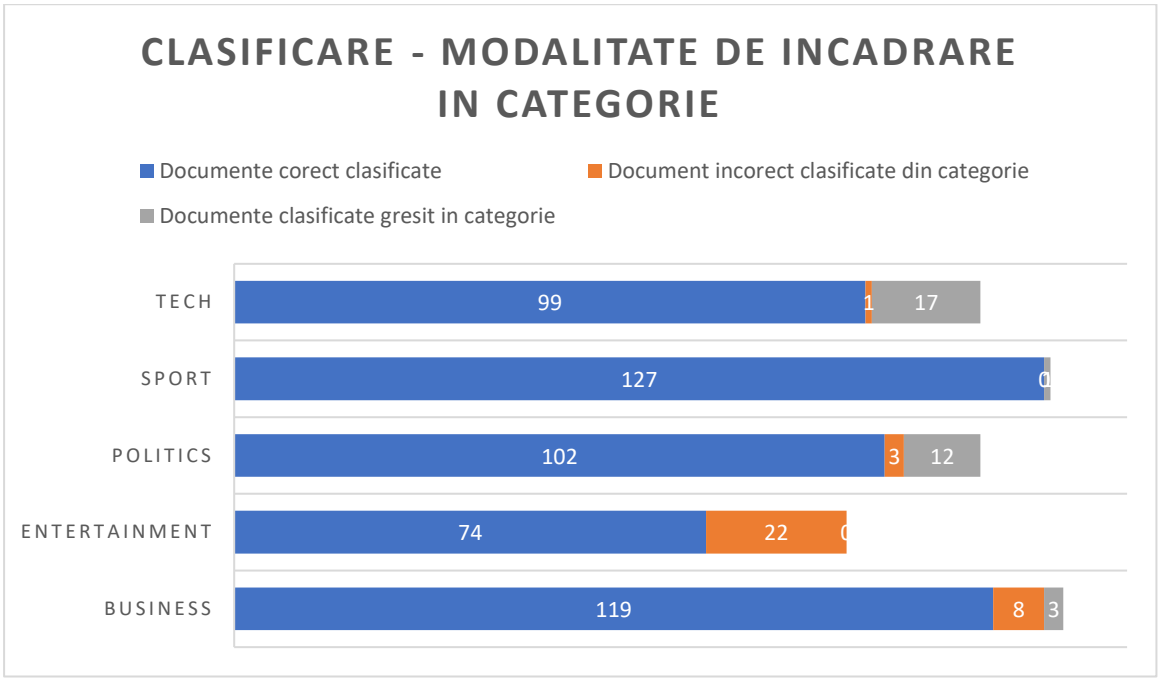
   Matricea de confuzie corespunzătoare:

	Business	Entertainment	Politics	Sport	Tech
Business	**119**	0	4	0	4
Entertainment	2	**74**	7	1	12
Politics	1	0	**102**	0	1
Sport	0	0	0	**127**	0
Tech	0	0	1	0	**99**

Procentajul de articole clasificate corect este de 94%, un procentaj ce poate fi considerat bun pentru algoritmul utilizat:



Clasificarea, pe categorii de articole, indică faptul că majoritatea categoriilor au înregistrat minimum un articol nu încadrat greșit în altă categorie. De asemenea, majoritatea articolelor care au fost clasificate greșit în această variantă au ajuns să fie clasificate ca fiind „tech”, următoarea opțiune fiind „politics”. Documentele din categorial „entertainment” au fost cel mai comun clasificate în altă categorie. De asemenea, toate articolele din categoriile „sport” și „tech” au fost încadrate în categoria corectă.



3. Varianta lematized, fără stop words

Valorile obținute au fost:

Precision: 0.9277

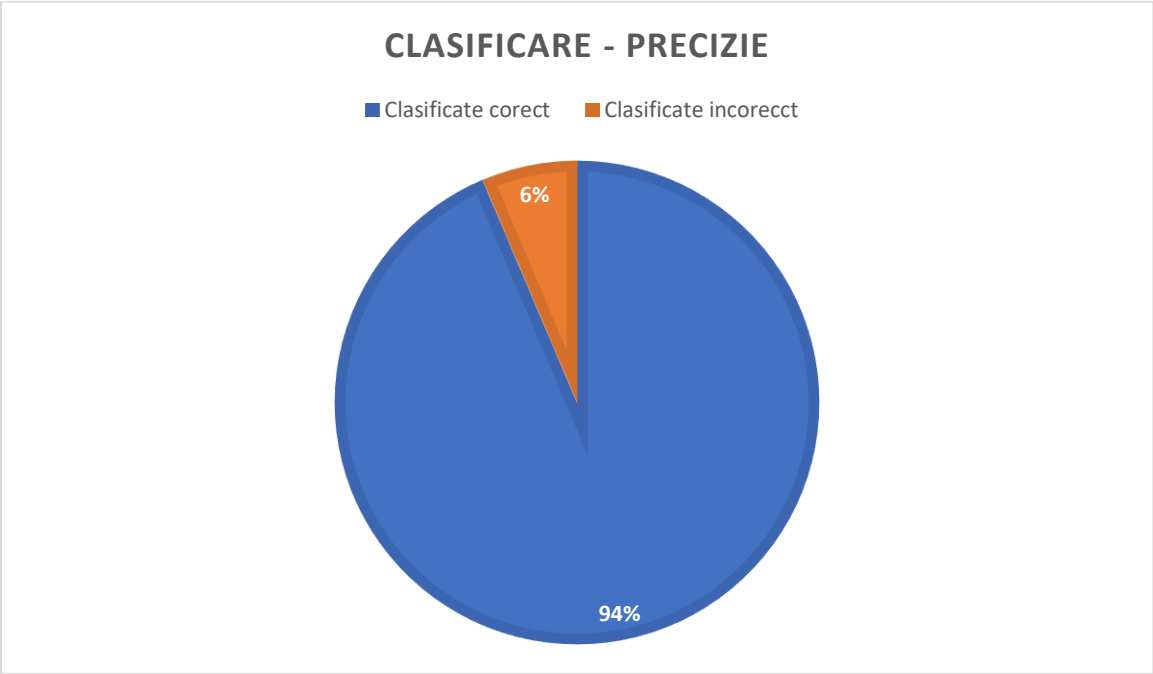
Recall :

- Business : 0.9291
- Entertainment : 0.6875
- Politics : 0.9903
- Sport : 1.0
- Tech : 1.0

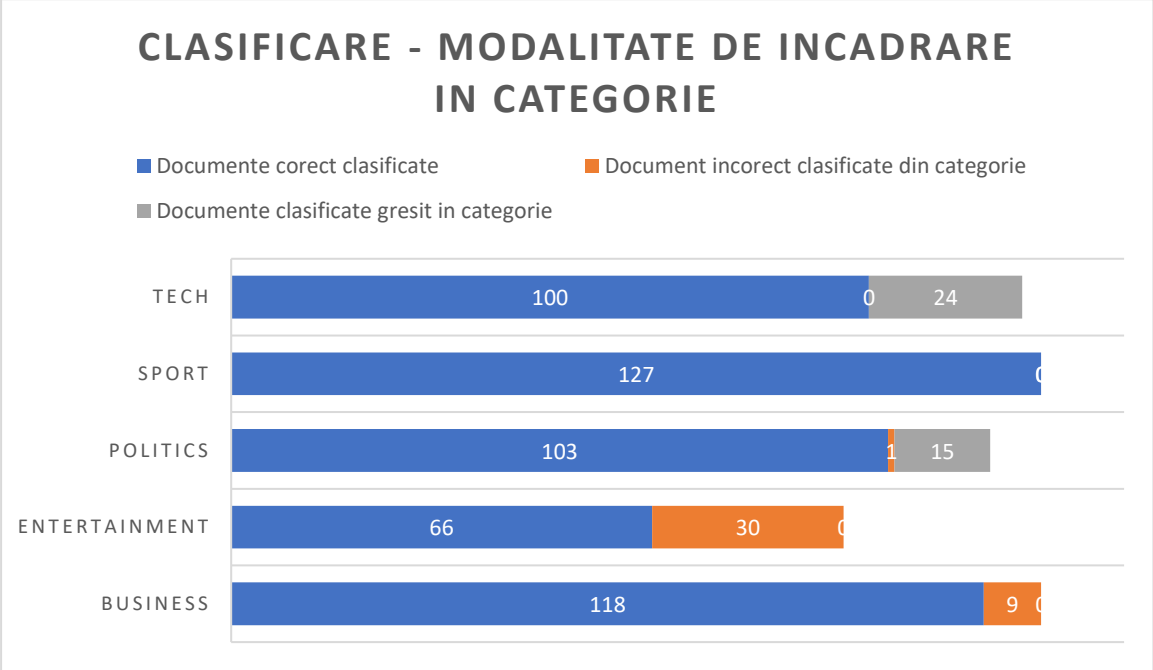
Matricea de confuzie corespunzătoare:

	Business	Entertainment	Politics	Sport	Tech
Business	**118**	0	6	0	3
Entertainment	0	**66**	9	0	21
Politics	0	0	**103**	0	1
Sport	0	0	0	**127**	0
Tech	0	0	0	0	**100**

Procentajul de articole clasificate corect este de 94%, un procentaj ce poate fi considerat bun pentru algoritmul utilizat.



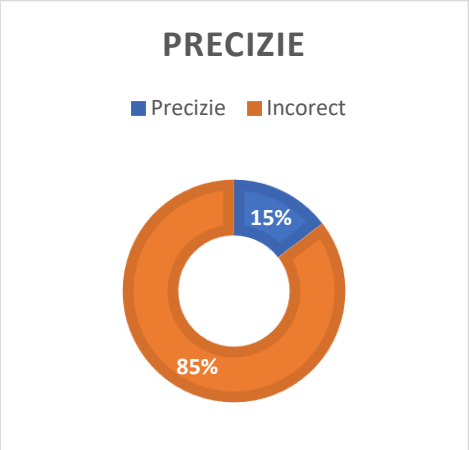
Clasificarea, pe categorii de articole, indică faptul că majoritatea categoriilor au înregistrat minimum un articol nu încadrat greșit în altă categorie. De asemenea, majoritatea articolelor care au fost clasificate greșit în această variantă au ajuns să fie clasificate ca fiind „tech”, următoarea opțiune fiind „politics”. Documentele din categorial „entertainment” au fost cel mai comun clasificate în altă categorie. De asemenea, toate articolele din categoriile „sport” și „tech” au fost încadrate în categoria corectă.

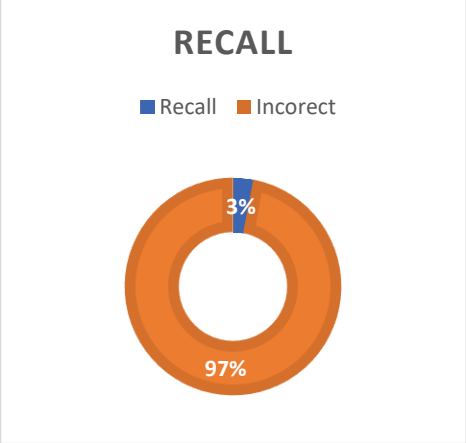


Sumarizare

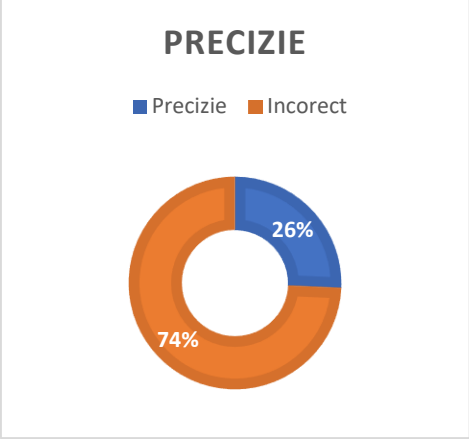
Pentru a testa capacitatea de clasificare a modelului, am rulat de mai multe ori, pentru fiecare variantă (unigrame/ bigrame) în parte de câte trei ori, pentru fiecare sub-cerință:

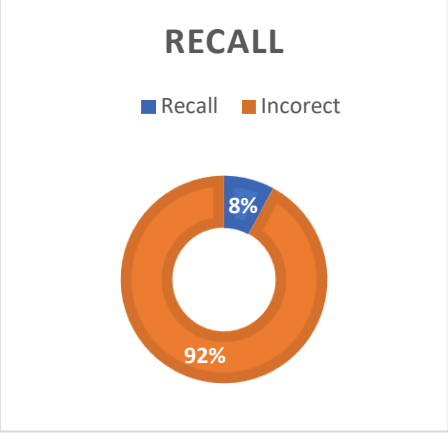
1. Unigrame
   a. Varianta tokenized, cu stop words incluse:
      - Precision: 0.1461
      - Recall: 0.0324



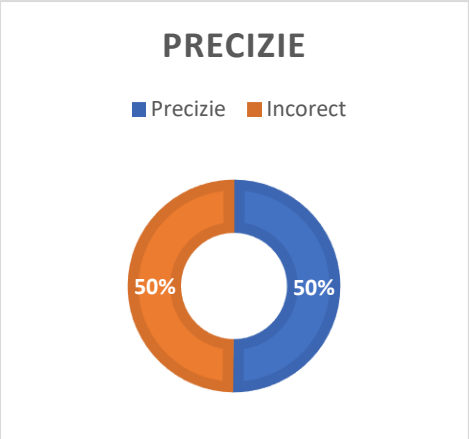


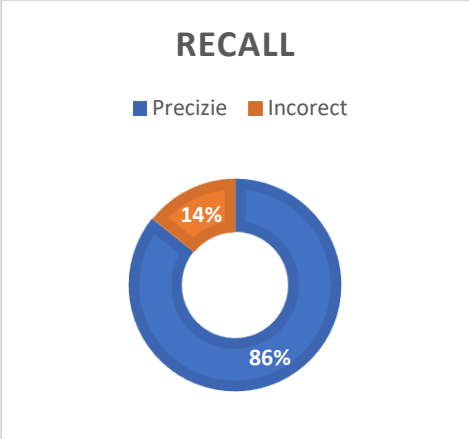
   b. Varianta tokenized, fără stop words :
      - Precision: 0.2570
      - Recall: 0.0789





c. Varianta lematized, fără stop words:
   - Precision: 0.5017
   - Recall: 0.8578





Singura variantă în care sumarizarea a înregistrat rezultate decente este cea în care sunt folosiți termeni lematizați, sumarizările fiind într-un procent de 50% similare cu cele corecte, iar recall-ul având un procent ridicat, de 85%.

2. Bigrame
   a. Varianta tokenized, cu stop words incluse:
      - Precision: 0.0
      - Recall: 0.0
   b. Varianta tokenized, fără stop words :
      - Precision: 0.0
      - Recall: 0.0
   c. Varianta lematized, fără stop words:
      - Precision: 0.0
      - Recall: 0.0

Pentru această variantă, nu am mai realizat grafice, întrucât rezultatele nu sunt satisfăcătoare.

5-fold cross validation

În continuare, pentru a putea realiza 5-fold cross validation, am împărțit setul de date în 5 părți egale pentru testare, urmând ca în antrenare să folosesc pentru fiecare parte cele 4 părți rămase.

Datele obținute în urma clasificării folosind 5-fold cross validation:

- Tokenized, cu stop words

	Precizie	Recall - business	Recall – entertainment	Recall -politics	Recall - sport	Recall - tech
Setul 1	0.9211	0.9313	0.6493	1.0	1.0	0.9875
Setul 2	0.9346	0.8823	0.7922	1.0	1.0	0.9875
Setul 3	0.8986	0.8921	0.6103	0.9705	0.9705	1.0
Setul 4	0.9099	0.8823	0.6493	0.9879	1.0	1.0
Setul 5	0.9420	0.9607	0.7179	1.0	1.0	1.0
Medie	0.9212	0.9097	0.6838	0.9916	0.9941	0.9950
Deviație standard	0.0158	0.0312	0.0643	0.0115	0.0118	0.0061

Valorile obținute sunt similare, deviația standard având o valoare foarte scăzută. Se poate concluziona faptul ca împărțirea datelor de test este realizată în mod aproximativ egal, astfel că nu există variații mari între cele 5 seturi testate.

- Tokenized, fără stop words

	Precizie	Recall - business	Recall – entertainment	Recall -politics	Recall - sport	Recall - tech
Setul 1	0.9617	0.9509	0.8701	0.9875	0.9901	1.0
Setul 2	0.9527	0.9411	0.8051	1.0	1.0	1.0
Setul 3	0.9594	0.9509	0.8701	0.9759	1.0	0.9875
Setul 4	0.9594	0.9411	0.8441	1.0	0.9901	0.9875
Setul 5	0.9420	0.9215	0.8076	0.9764	1.0	0.9876
Medie	0.9550	0.9411	0.8394	0.98796	0.9960	0.9925
Deviație standard	0.0071	0.0107	0.0286	0.0106	0.0048	0.0061

Valorile obținute sunt similare, deviația standard având o valoare foarte scăzută. Se poate concluziona faptul ca împărțirea datelor de test este realizată în mod aproximativ egal, astfel că nu există variații mari între cele 5 seturi testate.

- Lematized, fără stop words

	Precizie	Recall - business	Recall – entertainment	Recall -politics	Recall - sport	Recall - tech
Setul 1	0.9121	0.9117	0.6233	0.9879	1.0	1.0
Setul 2	0.9301	0.9607	0.6623	1.0	1.0	0.9875
Setul 3	0.9166	0.8921	0.6753	1.0	0.9901	1.0
Setul 4	0.9144	0.9019	0.6623	0.9879	1.0	0.9875
Setul 5	0.9064	0.8725	0.6666	1.0	0.9708	1.0
Medie	0.9159	0.90778	0.65796	0.99516	0.99218	0.995
Deviație standard	0.00786	0.0294	0.01796	0.00592	0.011356	0.006123

Valorile obținute sunt similare, deviația standard având o valoare foarte scăzută. Se poate concluziona faptul ca împărțirea datelor de test este realizată în mod aproximativ egal, astfel că nu există variații mari între cele 5 seturi testate.

Datele obținute în urma sumarizării folosind 5-fold cross validation, pentru unigrame:

- Tokenized, cu stop words

	Precizie	Recall
Setul 1	0.1242	0.0388
Setul 2	0.1381	0.0354
Setul 3	0.1164	0.0346
Setul 4	0.1211	0.0300
Setul 5	0.1211	0.0334
Medie	0.1241	0.0344
Deviație standard	0.0073	0.0028

Valorile obținute sunt similare, deviația standard având o valoare foarte scăzută. Se poate concluziona faptul ca împărțirea datelor de test este realizată în mod aproximativ egal, astfel că nu există variații mari între cele 5 seturi testate. Precizia și valoarea recall-ului sunt reduse, drept urmare cel mai probabil există o eroare pe care nu am reușit să o identific.

- Tokenized, fără stop words

	Precizie	Recall
Setul 1	0.1066	0.0331
Setul 2	0.1096	0.0386
Setul 3	0.1284	0.0312
Setul 4	0.1211	0.0305
Setul 5	0.1297	0.0362
Medie	0.1190	0.0339
Deviație standard	0.0094	0.0030

Valorile obținute sunt similare, deviația standard având o valoare foarte scăzută. Se poate concluziona faptul ca împărțirea datelor de test este realizată în mod aproximativ egal, astfel că nu există variații mari între cele 5 seturi testate. Precizia și valoarea recall-ului sunt reduse, drept urmare cel mai probabil există o eroare pe care nu am reușit să o identific.

- Lematized, fără stop words

	Precizie	Recall
Setul 1	0.4876	0.8413
Setul 2	0.4718	0.8396
Setul 3	0.4777	0.8534
Setul 4	0.4756	0.8666
Setul 5	0.4828	0.8328
Medie	0.4791	0.8467
Deviație standard	0.0055	0.0111

Valorile obținute sunt similare, deviația standard având o valoare foarte scăzută. Se poate concluziona faptul ca împărțirea datelor de test este realizată în mod aproximativ egal, astfel că nu există variații mari între cele 5 seturi testate. Valoarea preciziei și a recall-ului indică faptul că numai această variantă poate produce rezultate decente, însă totuși slabe.

Pentru datele obținute în urma sumarizării folosind 5-fold cross validation, pentru bigrame, nu am mai realizat tabele întrucât acestea ar fi avut numai valori de „0”.

Concluzii personale

Pentru clasificare, algoritmul obține rezultate bune, reușind cu o precizie de peste 90% (94% în majoritatea cazurilor) să încadreze articolele în categoriile corecte. Setul de antrenare conține 75% din totalul articolelor, suficient de mult încât să poată fi obținute probabilități cât de corecte posibil pentru cuvintele utilizate, ducând la o ușoară sortare a acestora în categorii. Cel mai bine sortate sunt articolele din zonele „politics", „sport" și „tech", domenii în care există o multitudine de cuvinte specifice lor, care fac facilă deciderea categoriei din care un articol face parte. Cel mai greu de sortat sunt articolele din zona „entertainment", în care limbajul folosit în general este unul clasic, lipsit de termeni specifici care să ajute la încadrarea acestora în această categorie.

Pentru sumarizare, algoritmul obține rezultate foarte slabe, în cazul unigramelor, iar pentru bigrame consideră toate propozițiile ca fiind irelevante, ceea ce duce la imposibilitatea realizării unor sumare. Singurul caz în care sumarizarea obține valori decente (medii) este în cazul în care cuvintele suferă procesul de lematizare, astfel că noile cuvinte rezultate pot avea o incidență mai mare în text, rezultând în probabilități de apariție mai mari, acest fapt reflectându-se în final în scorul înregistrat de propoziție ca fiind relevantă sau irelevantă. Algoritmul pare să nu funcționeze pe bigrame, lucrul ce nu ar trebui să se întâmple, drept urmare pot afirmă că undeva, în cadrul implementării, se află o eroare de logică pe care nu am reușit să o identific.

In cadrul „experimentului" ce folosește *„5 fold cross validation"*, valorile par să confirme ipotezele enunțate mai sus. Deviațiile standard au valori în general reduse (sub 0.01), iar valorile obținute sunt asemănătoare în fiecare set, astfel că modelul nu pare a fi under fit sau over fit în cazul clasificării. Rezultatele sunt în aceeași notă și în cazul sumarizării, însă problema rezultatelor slabe rămâne valabilă în cazul tuturor celor 5 iterații, de aici putând trage concluzia unei erori de logică.