UNIVERSITATEA DIN
BUCUREȘTI

FACULTATEA DE
MATEMATICĂ ȘI
INFORMATICĂ

SPECIALIZAREA INFORMATICĂ

Lucrare de licență

ANALIZA DIALECTELOR BRITANICE ALE LIMBII ENGLEZE

Absolvent
Cristiana Cocheci

Coordonator științific
Conf.dr. Sergiu Nisioi

București, iunie 2025

Rezumat

Această lucrare investighează posibilitatea clasificării automate a dialectelor limbii engleze scrise, în contextul formal al discursurilor parlamentare, concentrându-se pe varietățile britanică, irlandeză și scoțiană. Sunt comparate atât metode clasice de învățare automată — precum Naive Bayes, Logistic Regression și Support Vector Classifier — cât și rețele neuronale și modele moderne de tip transformer, precum BERT și ModernBERT. Analiza cantitativă arată că ModernBERT atinge cea mai ridicată acuratețe, dar metodele tradiționale rămân competitive, în special când sunt susținute de reprezentări bine alese ale textului (Tf-Idf, Bow), în timp ce sentence embedding-urile introduse în rețele neuronale sunt evaluate cel mai slab. În plus, prin utilizarea unor instrumente de interpretabilitate precum SHAP, Lime și Captum, este realizată o analiză calitativă detaliată, care scoate în evidență diferențe semnificative în alegerea cuvintelor, construcțiile sintactice și identitățile naționale al parlamentarilor din fiecare regiune.

Abstract

This paper explores the feasibility of automatically classifying written English dialects within the formal context of parliamentary speeches, focusing on British, Irish, and Scottish varieties. Both traditional machine learning methods — such as Naive Bayes, Logistic Regression, and Support Vector Classifier — and neural networks and modern transformer-based models like BERT and ModernBERT are compared. Quantitative analysis shows that ModernBERT achieves the highest accuracy, but traditional methods remain competitive, especially when supported by well-chosen text representations (Tf-Idf, BoW), while sentence embeddings used in neural networks yield the weakest results. Additionally, by employing interpretability tools such as SHAP, Lime, and Captum, a detailed qualitative analysis reveals significant differences in word choice, syntactic constructions, and the expression of national identity among parliamentarians from each region.

Cuprins

Capitolul 1

Introducere

Diversitatea lingvistică joacă un rol esențial în dezvoltarea competențelor Inteligenței Artificiale, în special în Procesarea Limbajului Natural. Beneficiile principale ale diversității lingvistice sunt: creșterea performanței modelelor de conversație și a mașinilor de traducere prin expunerea la o varietate mai mare ale aceleiași limbi, precum și îmbunătățirea uneltelor de antropologie lingvistică precum identificarea autorului unui text. Mai mult decât atât, rezultatele teoretice ne aduc mai aproape de înțelegerea evoluției limbajului uman și a relației dintre acesta și popoarele care îl ultilizează.

Lucrarea de față investighează posibilitatea clasificării automante a dialectelor limbii engleze în contexte formale scrise, mai precis dintre britanică, irlandeză și scoțiană. Se pornește de la ipoteza că, deși forma scrisă formală a limbii standardizează exprimarea, pot exista diferențe subtile în alegerea cuvintelor, construcțiile sintactice și stilul discursiv care reflectă originea dialectală a vorbitorului.

Contribuții personale

În această lucrare demonstrăm că, deși ModernBERT obține cele mai bune rezultate în clasificarea dialectelor și oferă îmbunătățiri semnificative față de versiunea clasică BERT, metodele tradiționale, precum clasificatorii probabilistici și Support Vector Classifier, rămân surprinzător de competitive și relevante. Pentru obținerea celor mai bune rezultate, am efectuat o analiză comparativă a metodelor de preprocesare a textului și a alegerii reprezentărilor sale. Mai mult decât atât, analiza calitativă nu doar validează performanțele tehnice ale modelelor, ci scoate la iveală diferențe subtile de sintaxă, lexic și stil discursiv între cele trei grupuri studiate. Astfel, lucrarea oferă atât contribuții tehnice valoroase în domeniul NLP, cât și perspective originale asupra modului în care identitatea națională și atitudinea politică transpar în discursul scris al parlamentarilor europeni.

În comparație cu lucrările existente în domeniu [1] [2] [3], care se concentrează preponderent pe limbaj vorbit sau pe corpusuri informale (de exemplu, social media), abordarea noastră se concentrează exclusiv pe texte formale scrise, într-un context politic (discursuri parlamentare), ceea ce aduce o contribuție originală în înțelegerea manifestării dialectelor

în registrul standardizat al limbii.

Structura lucrării

Lucrarea este structurată astfel: începem prin prezentarea metodologiei, setul de date, și clasificatorii utilizați, împreună cu detalii despre preprocesare și extragerea trăsăturilor. Urmează analiza comparativă a rezultatelor și o discuție calitativă asupra predicțiilor, inclusiv interpretabilitatea modelelor transformer. În încheiere, sunt discutate limitările lucrării și sunt propuse direcții viitoare de cercetare.

Capitolul 2

Metode utilizate

2.1 Setul de Date

Pentru antrenarea și evaluarea modelelor am folosit setul de date Europarl (European Parliament Proceedings Parallel Corpus) [4], care conține discursuri originale din Parlamentul European din perioada 1996-2011, precum și traducerile lor autorizate și aliniate. Textele se regăsesc în: engleză, germană, olandeză, daneză, franceză, italiană și spaniolă.

Din acesta am extras discursuri ale parlamentarilor britanici, irlandezi și scoțieni în varianta lor originală, adică în limba engleză. După ce am filtrat din date exemplele prea scurte sau invalide, am construit un nou set de date echilibrat cu cele trei tipuri de discursuri (toate în limba engleză) de 6060 de exemple - 2020 din fiecare categorie.

Set de date	Limbi	Cuvinte	Cuvinte unice	Medie cuv/txt	Texte brit.	Texte irland.	Texte scoț.	Total texte
Înainte	7	3.994.009	86.571	259	6053	2810	2223	11.086
După	1-eng.	1.536.585	64.066	254	2020	2020	2020	6060

2.2 Clasificare prin Modele Probabilistice și SVM

Primul set de modele utilizate a fost cel probabiilistic, incluzând:

- **Naive Bayes [5]**: Naive Bayes este o clasă de clasificatori probabilistici care presupun că feature-urile sunt condițional independente, dată fiind clasa target. Natura nerealistă a acestei presupuneri, numită independență naivă, este ceea ce i-a câștigat numele. Dezavantajul Naive Bayes este că adesea produce probabilități exagerat de încrezătoare. Pe de altă parte, clasificatorul este foarte scalabil, având nevoie de un singur parametru per feature.

  Naive Bayes este un **model generativ** deoarece utilizează probabilitatea condiționată $P(x|c)$ (likelihood), care descrie modul în care sunt generate caracteristicile (features) unui document *presupunând că acesta aparține clasei c.* [6]

- **Bernoulli NB** [7]: Este bazat pe date binare: fiecare termen (cuvânt sau Ngram) din vectorul de feature-uri este asociat cu valoarea 1 sau 0 (1 - cuvântul apare în document, 0 - nu apare). Probabilitatea unei clase c dat un document $x = (w_1...w_n)$ poate fi scrisă ca:
  $P(c|x) = P(c) \prod_{i=1}^{n} P(w_i|c)^b (1 - P(w_i|c))^{(1-b)}$, unde b $\in 0, 1$.
- **Multinomial NB** [7]: Aceasta ia în considerare și frecvența termenilor. Probabilitatea unei clase c dat un text $x = (w_1...w_n)$ și x_i numărul de apariții al lui w_i în text, poate fi scrisă ca:
  $P(c|x) = P(c) \prod_{i=1}^{n} P(w_i|c)^{x_i}$
  Putem observa următoarea diferență între cele două metode: Bernoulli modelează prezența *și* absența cuvintelor, nu frecvența sau importanța lor.

- **Regresia Logistică** [6]: Pentru clasificarea a trei clase, folosim Regresia Logistică Multinomială, numită și *softmax regression*. Spre deosebire de Naive Bayes, aceasta este un **model discriminativ**, deoarece încearcă să calculeze direct $P(x|c)$. Astfel învață să atribuie ponderi mai mari unui feature care ajută direct la discriminarea dintre clasele posibile, chiar dacă nu poate genera un exemplu dintr-o clasă.

  Regresia Logistică are următorul avantaj față de Naive Bayes: este mult mai robustă la feature-uri corelate. Dacă un feature f ar apărea de două ori, ca f_1 și f_2 , NB le-ar considera diferite și ar multiplica probabilitățile lor, supraestimând încrederea în predicție, dar LR ar atribui o parte din pondere la w_1 și o parte la w_2, având un rezultat mai realist.

- **SVC (Support Vector Classifier)** [8]: O metodă de învățare supervizată. Poate fi aplicată pe date neliniare, iar decision boundary-ul este tot neliniar. Folosește hiperplanul de margine maximală pentru a alege un decision boundary. Implementarea aleasă [9] utilizează kernel-ul RBF (Radial Basis Function), $C = 1.0$, $gamma = scale$ și metoda de clasificare *one-versus-one* pentru cele 3 clase. Practic se construiește un SVM pentru fiecare pereche de clase (în cazul nostru 3 SVM-uri). Cele 3 SVM-uri sunt aplicate pe noul sample, iar rezultatul cel mai frecvent este ales ca predicție.

Pentru aceste clasificatoare am folosit următoarele feature-uri:

- **CountVectorizer (Bag Of Words)** : textele sunt convertite în vectori de lungime fixă în care se numără aparițiile fiecărui cuvânt.

- **CountVectorizer + NGrame (1 - 3 cuvinte)** : de data aceasta textul nu se sparge în cuvinte unice, ci în secvențe de 1, 2 sau 3 cuvinte adiacente. Vectorul se construiește identic cu punctul anterior.

- **TF-IDF** : reprezentarea este similară cu cea de la Countvectorizer, dar ia în considerare și importanța cuvântului pentru textul respectiv. Astfel se calculează frecvența fiecărui cuvânt în text (Term frequency) și inversul frecvenței documentelor ce conțin cuvântul respectiv (Inverse Document Frequency).

- **TF-IDF + NGrame (1-3 cuvinte)** : idem. + Ngrame.

K-fold Cross Validation

Având un set de date de dimensiuni reduse am ales să folosesc ca metodă de resampling K-Fold Cross Validation (K=10) [10]. Setul de date este împărțit în 10 segmente de dimensiuni egale. La fiecare iterație de antrenare, sau Fold, sunt alese 9 segmente pentru antrenare, iar al 10-lea pentru validare. K-fold cross-validation ajută ca rezultatele să fie mai putin optimiste (sau biased), cum ar apărea în urma folosirea unui simplu train-test split.

Feature Selection

Alegerea feature-urilor, sau reducerea dimensionalității, se folosesc atât pentru îmbunătățirea performanței modelelor, cât și pentru creșterea acurateței lor.

Pentru alegerea feature-urilor am folosit $SelectKBest$, cu $K = 10000$, din modulul sklearn.feature_selection [11]. $SelectKBest$ folosește metode de statistică univariată pentru a selecta primele K cele mai performante feature-uri. Funcția de evaluare folosită a fost $chi2$ (χ^2). Aceasta testează independența între fiecare feature în parte si clasă. Funcția este folosită preponderent în probleme de clasificare care nu au feature-uri negative (precum bag-of-words sau TF-IDF, unde feature-urile reprezintă frecvențele termenilor.

2.3 Rețele Neuronale

2.3.1 Sentence embbedings

Sentence embbeding-urile [12] transpun propozițiile în vectori denși de lungime fixă. Embedding-urile universale sunt pre-antrenate pe seturi mari de date și pot fi utilizate cu succes în o multitune de downstream-tasks (analiza sentimentelor, clasificare, traducere). Acestea sunt o formă de **transfer-learning**, paradigmă ce s-a dovedit indispensabilă în NLP-ul modern.

În principal, Sentence embedding-urile sunt folosite pentru capacitatea lor de a reprezenta înțelesul semantic, dar studii precum [13] arată că acestea capturează și caracteristici stilistice, precum identificarea unui autor. Cele mai bune rezultate ale studiului menționat le-a avut **all-MiniLM-L12-v2** cu F1 score de 0.623, care a fost leader-ul clasamentului pe setul de testare PAN23 de verificare a autorilor.

În cazul textului scris, există o suprapunere între detectarea autorilor și a dialectului. Acestea au în comun diferențele lingvistice subtile, sunt independente de mesajul textului,

pot fi modelate folosind modele de distribuție sau modele de embedding-uri etc. Diferența dintre cele două este că în timp ce problema de identificare a autorului caută idiosincrasie, cea a dialectului analizează variația sistematică împărtășită de un grup. Considerând punctele comune, am ales să le încercăm și pentru deosebirea nuanțelor lingvistice ale unui dialect.

2.3.2 Arhitectura rețelei

A doua metodă de clasificare aleasă a fost o rețea neuronală adâncă. Aceasta a primit trei tipuri de input :

1. **sentence embedding**-uri extrase cu ajutorul librăriei SentenceTransformers [14] cu **all-MiniLM-L6-v2**. Modelul trece textul printr-un tokenizer, apoi printr-un transformer cu 6 straturi. Fiecare token primește un embedding contextual din stratul final. Peste aceste embeddinguri se aplică operația de mean-pooling.

2. token-ul **CLS** din **BERT**

3. token-ul **CLS** din **ModernBERT** [15]. Atât la BERT cât și la ModernBERT, token-ul CLS reprezintă o sumarizare a textului introdus. Este un embedding calculat separat ce nu aparține unui token, ci întregului text.

Ca *Loss Function* am folosit Cross Entropy, iar ca optimizator Adam [16]. Pentru a preveni overfitting-ul am utilizat următoarele 5 metode:

1. **dropout** (la forward pass printr fiecare strat un subset aleator de 50% din numărul total de neuroni este dezactivat),

2. **batch normalization** [17] (un strat neural cu doi parametrii antrenabili β și γ care normalizează, scalează și translatează output-ul fiecărui strat, făcând backpropagation-ul mai lin și eficient),

3. **early stopping** (alegerea unei iterații din mijlocul antrenării modelului, pentru a păstrarea parametrilor care avantajează mai mult o metrică a validării decât una a antrenării),

4. **weight decay**, sau regularizare L2 (penalizează weight-urile mari) și

5. **label smoothing** [18] (folosirea așa numitelor *soft labels*, adică înlocuirea vectorilor de probabilități one-hot cu media ponderată dintre label-uri și distibuția lor uniformă).

Arhitectura acesteia este următoarea: inputul este un vector de Embedding-uri de dimensiune 384 pentru Sentence Transformer(*all-MiniLM-L6-v2*) și 768 pentru CLS tokens

(BERT și ModernBERT). Urmează 5 Straturi Fully connected: 400, 300, 200, 100, 50 de neuroni. Dupa fiecare strat aunt aplicate: Batch Normalization, LeakyReLu (funcția de activare) și Dropout. Output : 3 valori reprezentând probabilitățile celor 3 clase. În cazul rețelei pentru ModernBERT CLS, straturile de neuroni a fost modificate la următoarele dimensiuni: 500, 400, 400, 300, 100.

2.4 Transformers

BERT [19], prescurtare pentru Bidirectional Encoder Representations from Transformers, este un model de Machine Learning pentru NLP de 110 milioane parameteri. A fost dezvoltat în 2018 de cercetătorii de la Google AI Language și a ajuns o soluție State Of The Art pentru majoritatea problemelor comune de Procesare a Limbajului Natural, precum clasificarea sentimentelor și named entity recognition.

BERT a fost antrenat pe datele Wikipedia (2.5B cuvinte) și Google's BooksCorpus (800M cuvinte). În total 3.3 Miliarde de cuvinte, 64 TPUs (Tensor Processing Units - acceleratoare de AI construite de Google, optimizate pentru antrenarea și inferența modelelor de AI) și 4 zile de antrenare continuă au dus la succesul lui BERT.

Șase ani mai târziu, în decembrie 2024, cercetătorii de la Answer.AI și LightOn au lansat **ModernBERT** [20]. ModernBERT este un model nou cu 149 milioane parametri, antrenat pe 2 Trilioane de cuvinte, și reprezintă o îmbunătățire a lui BERT atât din punct de vedere al vitezei cât și al eficienței [21]. Acesta preia mai multe optimizări ale LLM-urilor din ultimii ani și le aplică pe un model tip BERT, căruia îi sunt adăugate schimbări mici de arhitectură și ajustări ale procesului de antrenare.

A treia metodă de clasificare aleasă este utilizarea celor două modele preantrenate BERT și ModernBERT. Acestea au fost importate din biblioteca Transformers de pe Hugging Face [22] ca "bert-base-uncased" și "answerdotai/ModernBERT-base" și finetuneate complet (fără înghețarea vreunui straturi) pe setul nostru de date. Textul a fost procesat prin tokenizer-ele corespunzătoare din aceeași bibliotecă.

2.5 Tehnologii de interpretare a rezultatelor

Pentru vizualizarea celor mai importante trăsături în efectuarea predicțiilor am utilizat următoarele bibliioteci de interpretabilitate: SHAP [23], Lime [24] și Captum [25]. SHAP (SHapley Additive exPlanations) explică predicțiile unui model calculând contribuția fiecărui feature la predicție pe baza valorilor Shapley din teoria jocurilor. Lime (Local Interpretable Model-agnostic Explanations) este o tehnică de aproximare a oricărui model de învățare automată cu un model local, interpetabil pentru fiecare predicție în parte. Captum ('înțelegere' în Latină) este o bibliotecă PyTorch care oferă diverse metode de interpretabilitate pentru a evidenția importanța feature-urilor în rețele neuronale.

Capitolul 3

Rezultate

3.1 Clasificare prin Modele Probabilistice și SVM

Pentru toate tipurile de feature-uri am aplicat KBestFeatureSelection, deci spațiul de reprezentare este mereu un vector de dimensiune 10000. Textele au fost preprocesate prin lematizare și scrise cu litere mici. Metrica folosită este acuratețea, deoarece clasele sunt perfect balansate, iar aceasta surprinde realist performanța modelelor. Pentru graficele următoare acuratețea este considerată media tuturor celor 10 folduri. În 3.3 este afișată în plus și deviația standard a modelelor.

3.1.1 Analiza comparativă a rezultatelor modelelor

În graficul din stânga figurii 3.1 avem rezultatele celor 3 modele probabilistice : regresia Logistică, Multinomial Naive Bayes și Bernoulli Naive Bayes și ale SVC, evaluate pentru fiecare dintre cele 4 seturi de feature-uri : Tf-Idf Unigrame, Tf-Idf Ngrame, BoW Unigrame și BoW Ngrame.

Presupuneri inițiale

Datele textuale sunt de obicei separabile liniar atunci când sunt reprezentate vectorial într-un spațiu multidimensional suficient de informativ [26]. Cum SVM-urile și Regresia Logistică încearcă să găsească o combinație liniară care să separe clasele, ne așteptam ca ele să aibă cele mai bune rezultate.

Performanțe maxime

Performanța maximă este cea a SVC, cu 0.6921 acuratețe. Aceasta este atinsă pentru două seturi de feature-uri: Tf-Idf cu Ngrame și BoW cu Ngrame. Putem deduce astfel că cel mai probabil spațiile vectoriale ale celor două reprezentări sunt similare dpdv geometric.

În grafice este ilustrat SVC cu RBF kernel. Acesta depășește, cu o diferențe mică, LinearSVC (cu kernel liniar), sugerând că datele nu sunt totuși complet liniar separabile în problema noastră.

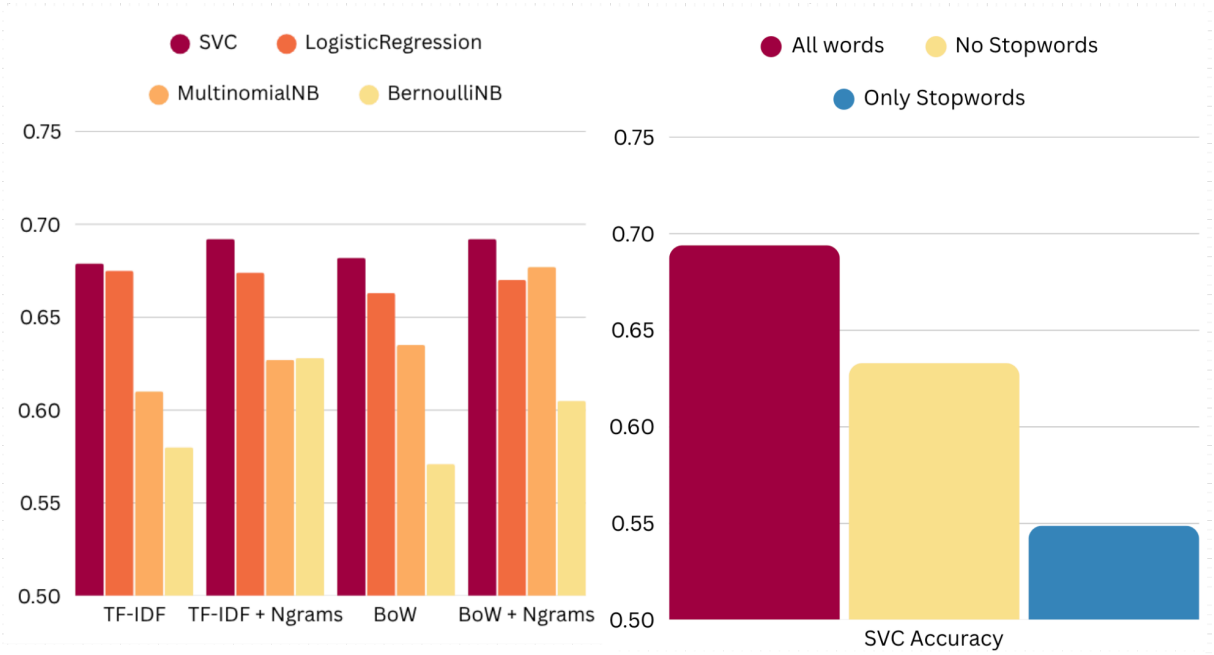

Figure 3.1.1 **Figure 3.1.2**

Figura 3.1: **3.1.1:** Acuratețea modelelor probabilistice și a SVC pe cele patru seturi de feature-uri. **3.1.2:** Acuratețea SVC pe trei tipuri de procesări ale textului.

Regresia Logistică, ce feature-uri o avantajează?

Regresia Logistică este consecventă pe toate feature-urile, cu un maxim de 0.675 acuratețe cu Tf-Idf pe cuvinte simple, la o diferență foarte mică de Tf-Idf cu Ngrame. Observăm cum capacitatea Regresiei Logistice de a trata valori reale date de Tf-Idf - nu binare, și de a modela interacțiunea dintre cuvinte prin atribuirea de ponderi crește acuratețea prezicerilor. În plus, utilizarea regularizării previne overfitting-ul și face clasificatorul mai robust la spații de dimensiuni mari, ca cel rezultat din metodele de reprezentare utilizate.

Multinomial Naive Bayes

Cum NB modelează prezența și absența cuvintelor, rezultatul pentru primele două seturi de feature-uri este explicabil. Acuratețea sa pentru BoW cu unigrame stagnează, dar pentru BoW cu Ngrame urcă semnificativ. Deducem că lipsa absolută a contorizării relațiilor dintre cuvinte scade rezultatele, deci pentru identificarea dialectelor este foarte importantă sintaxa propozițiilor.

MultinomialNB vs. Regresia Logistică

Pentru Tf-Idf, NB are rezultate considerabil mai slabe decât Regresia Logistică. Diferența se micșorează pentru BoW, iar pentru BoW cu Ngrame, NB reușește să depășească Regresia Logistică. Această performanță este cel mai probabil datorată setului de date mic zgomotos. Mai mult, cuvintele pe care se bazează deosebirea claselor sunt unele comune, deci cuantificarea lor binară nu este suficient de reprezentativă în distingerea între clase.

BernoulliNB vs. MultinomialNB

MultinomialNB are performanțe mai bune ca BernoulliNB în 3 din 4 cazuri evaluate,

așa cum literatura ne sugerează [27]. Câteva motive sunt că diferențele între cele trei dialecte în engleza formală în general nu sunt determinate de prezența sau absența cuvintelor, ci mai degrabă de frecvența utilizării lor, sau de faptul că MultinomialNB scalează mai bine cu dimensiunea textelor și a vocabularului. În plus, pentru dialecte, absența anumitor cuvinte comune nu este neapărat semnificativă, deci cuantificarea lor de către BernoulliNB poate fi un factor negativ.

În cea de-a doua coloană de feature-uri (Tf-Idf + 1-3 Ngrams), observăm că MultinomialNB și BernoulliNB au performanțe egale, spre deosebire de celelalte coloane. Acest lucru este de așteptat, deoarece Ngramele fac numărul feature-urilor să crească, iar frecvența fiecărei Ngrame în parte să scadă. Practic, se introduce un număr foarte mare de combinații posibile de cuvinte care apar de puține ori. Așadar, vectorii Tf-Idf devin sparse, comportându-se similar cu unii binari, la care Bernoulli s-ar aștepta, lucru care duce la creșterea performanței sale.

3.1.2 Evaluarea tipurilor de procesare a textului

Presupuneri initiale

Așa cum s-a demonstrat în analiza de mai sus, natura problemei de clasificare face sintaxa propozițiilor mai importantă decât sensul acestora. Din aceste considerente, am comparat trei tipuri de preprocesare de text: păstrând toate cuvintele, eliminând cuvintele de legătura sau folosind doar cuvinte de legătură (figura 3.1.2). Am presupus că doar testarea cuvintelor de legătură ar trebui să fie suficientă pentru o clasificare brută a dialectelor și că eliminarea lor va avea rezultate mai slabe.

Interpretarea diferentelor

- **All words - 0.67**. Includerea tuturor cuvintelor are cea mai bună performanță, ceea ce sugerează două posibilități: atât cuvintele de conținut cât și cele de legătură semnalează diferențe de dialect, sau clasificatoarele se bazează și pe diferențe de conținut în evaluare. Răspunsul îl primim în urma analizei calitative din următorul capitol. Ambele variante sunt adevărate într-o măsură, dar clasificatoarele reușesc majoritar să facă preziceri independente de topicul textului. În orice caz, astfel se capturează un spectru lingvistic mai variat (alegerea vocabularului, sintaxă, frazare).

- **No stopwords - 0.65**. Eliminarea cuvintelor de legătură reduce performanța cu puțin, ceea ce înseamnă că deosebirea se bazează majoritar pe cuvinte de conținut.

- **Only stopwords - 0.54**. Atribuirea aleatorie a label-urilor rezultă într-un scor de 0.33, dar rezultatul obținut în urma evaluării textelor ce conțin numai cuvinte de legătură este semnificativ peste scorul de bază. Așadar, frecvența și ordinea cuvintelor funcționale diferă între dialecte. Această observație este susținută de

surse sociolingvistice [28][29]: dialectele adeseori variază subtil în sintaxă, folosirea verbelor auxiliare, construcția negației, alegerea prepozițiilor. Cu toate acestea, studiile pun în lumină în special diferențele din limbajul colocvial și preponderent verbal. Multe dintre aceste subtilități sunt pierdute în limbajul formal scris.

Figure 3.2.1 **Figure 3.2.2**

Figura 3.2: **3.2.1:** Acuratețea rețelelor neuronale pentru cele trei tipuri de feature-uri. **3.2.2:** Acuratețea BERT și ModernBERT pentru două tipuri de text de intrare.

3.2 Rețele Neuronale

În stânga figurii 3.2 sunt prezentate rezultatele clasificării cu ajutorul unei rețele neuronale cu 5 straturi ascunse cu embedding-uri extrase din all-MiniLM-L6-v2, BERT și ModernBERT. Arhitectura rețelei poate fi vizualizată în imaginea 6.1.

Rezultate maxime

Cea mai mare acuratețe obținută este de 0.64 în urma utilizării token-ului CLS din BERT. Reprezentarea CLS capturează semnale mai bune de diferențiere între dialecte, probabil datorită atenției către anumite poziții ale cuvintelor sau sintaxei. Curbele de antrenare ale modelului pot fi vizualizate în imaginea 6.2.

BERT CLS vs. Sentence embeddings - all-MiniLM-L6-v2

Cele două tipuri de embeddinguri sunt calculate în moduri diferite și au dimensiuni diferite, cele CLS fiind duble (768) față de cele all-MiniLM-L6-v2 (384). Dimensiunea mai mare permite încorporarea mai multor informații despre propoziție, în special de sintaxă, care poate nu sunt prioritare în reprezentarea mai redusă, unde sensul textului ar fi de interes.

ModernBERT CLS

Deși ModernBERT este un model îmbunătățit, rețeaua care folosește token-ul CLS are cea mai slabă acuratețe. Posibil, embedding-urile CLS nu sunt bine optimizate pentru

clasificare sau au nevoie de finetuning pentru a extrage feature-uri utile. Acuratețea a crescut odată cu numărul de neuroni, sugerând că reprezentările sunt mai complexe și mai puțin discriminative în mod direct și necesită un clasificator mai puternic. Totuși, creșterea excesivă a dus la overfitting, iar performanța finală a fost nesatisfăcătoare. De vină este cel mai probabil dimensiunea redusă a setului de date în comparație cu complexitatea modelului.

3.3 Transformers

Graficul din dreapta figurii 3.2 ilustrează performața BERT și ModernBERT finetuned pe setul nostru de date timp de câte 3 epoci. Vedem în coloana stângă performanța lor pe textul original, neprocesat, iar în coloana dreaptă asupra textului au fost aplicate următoarele modificări: eliminarea semnelor de punctuație, scrierea cuvintelor cu litere mici.

Diferența între cele două tipuri de date de intrare

Eliminarea punctuației și aducerea literelor în forma minusculă reduc în cazul ambelor modele performanța cu 1-2 procente. Scăderea sugerează că și capitalizarea și punctuația au roluri subtile în transmiterea informațiilor de sens și dialect, posibil prin structura propozițiilor, stilul de abrevieri obiceiuri de capitalizare. Aceste trăsături sunt, în orice caz, auxiliare, impactul lor fiind unul puțin semnificativ.

Performanța ModernBERT vs. BERT

Modelul ModernBERT prezintă performanțe superioare în comparație cu predecesorul său. Acesta beneficiază de un set de antrenament extins, de aproximativ 1000 de ori mai mare decât cel utilizat pentru BERT, oferind o acoperire mai amplă a variațiilor dialectale. De asemenea, ModernBERT integrează un tokenizator de generație nouă, care s-a dovedit considerabil mai robust în fața cuvintelor necunoscute, cum ar fi cele rezultate din erori ortografice. În contexte colocviale, această caracteristică i-ar conferi un avantaj în identificarea termenilor specifici anumitor dialecte. Cu toate acestea, corpusul utilizat în evaluare constă în enunțuri formale și standardizate, ceea ce limitează relevanța acestor diferențe în practică. Astfel, performanțele celor două modele rămân comparabile, iar concluzia este că atât BERT, cât și ModernBERT demonstrează o capacitate similară de procesare a textelor în contextul dat.

3.4 Comparație finală. Media și deviația standard a acurateței.

În graficul de mai jos, figura 3.3, sunt afișate cele mai bune rezultate de acurateță ale fiecărui model testat pe feature-urile cele mai avantajoase. Bara corespunzătoare

fiecărui model reprezintă deviația lor standard pentru evaluarea pe cele 10 folduri din Cross Validation.

Rezultatele prezentate indică o competiție strânsă între cele mai performante modele, cu ModernBERT obținând cea mai mare acuratețe, dar cu o marjă de eroare mai mare, urmat îndeaproape de SVC. Diferența dintre cele două este minimă, sugerând că, deși ModernBERT beneficiază de o arhitectură avansată și de un set de antrenament extins, metodele clasice bine calibrate — precum SVC combinat cu reprezentări Tf-Idf — pot concura eficient cu modelele de tip transformer în sarcini de clasificare bine definite. Performanța BERT completează topul clasificatoarelor testate.

În contrast, rețelele neuronale alimentate cu reprezentări simplificate obțin rezultate mai slabe, ceea ce sugerează că, în absența unei fine-tunări adaptate, aceste reprezentări nu sunt suficient de expresive pentru diferențierea dialectală.

În concluzie, deși metodele moderne aduc îmbunătățiri, utilitatea modelelor clasice rămâne relevantă, iar selecția optimă depinde de contextul aplicației și de resursele disponibile.

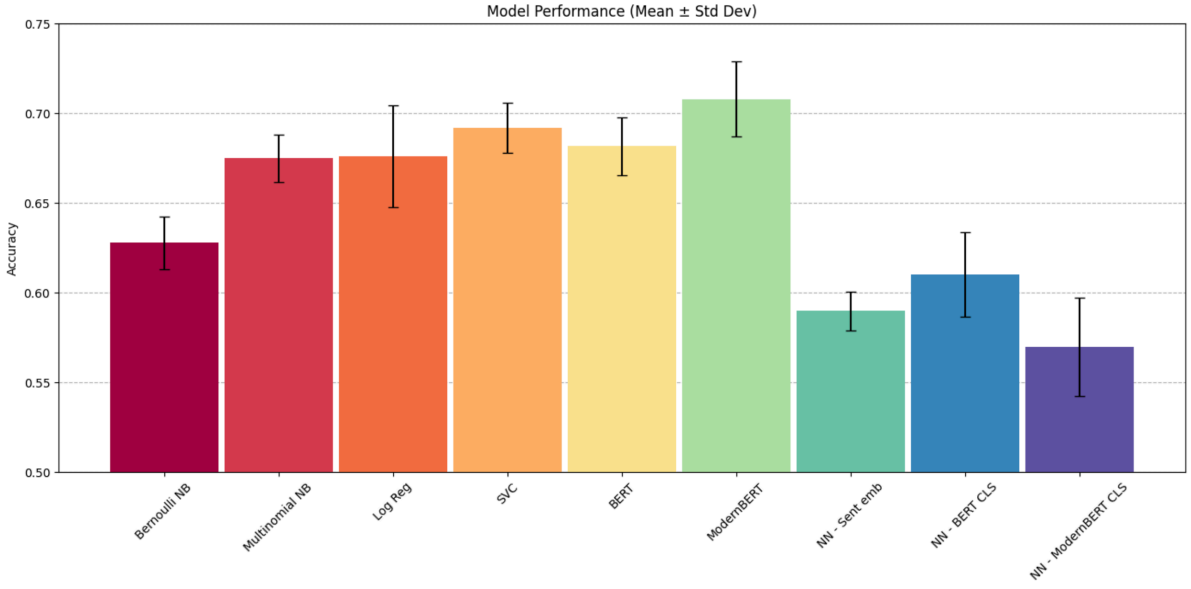

Figura 3.3: Comparație a performanței tuturor modelelor.

Capitolul 4

Analiza Calitativă

4.1 Clasificare prin Modele Probabilistice

Scopul analizei globale SHAP este de a descrie comportamentul statistic anticipat al unui model referitor la întreaga distribuție a valorilor feature-urilor din input. Acest lucru se obține prin agregarea valorilor SHAP pentru instanțe individuale din întreaga populație. În figura 4.1 avem un astfel de rezumat global pentru regresia logistică.

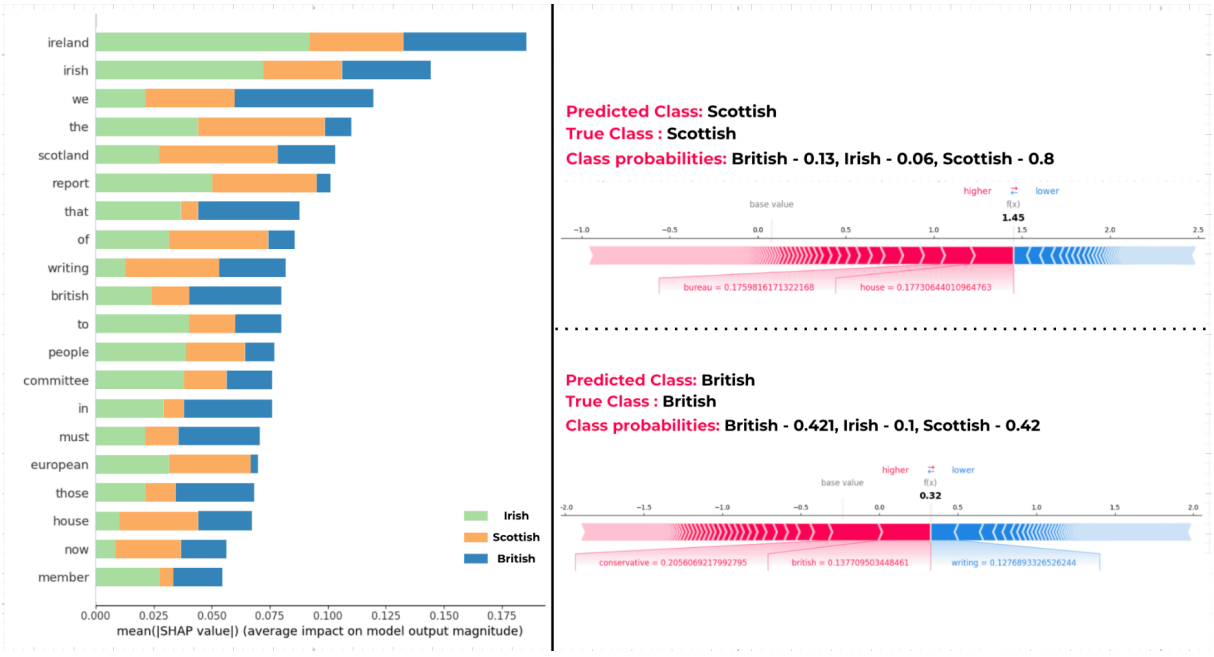

Figura 4.1: **Stânga**: Analiza globală SHAP pentru regresia logistică. **Dreapta**: Vizualizarea predicțiilor a două texte individuale cu SHAP.

Consultarea matricelor de confuzie, prezentate în Anexă la 6.3, indică în mod constant că modelele clasificatoare identifică cel mai precis textele redactate în dialectul irlandez. Conform datelor din analiza globală SHAP, cele mai relevante două trăsături predictive sunt termenii 'Ireland' și 'irish'. Alte trăsături cu valoare predictivă semnificativă, precum

'Scotland' și 'british', sunt asociate în mod corespunzător cu celelalte dialecte. Aceastea sugerează că parlamentarii își menționează frecvent propria identitate națională, ceea ce facilitează recunoașterea automată a textelor. Observațiile oferă mai degrabă informații de natură tematică decât lingvistică, relevând totuși o posibilă tendință a parlamentarilor irlandezi de a-și afirma identitatea mai pronunțat.

Din analiza golbală SHAP mai reiese că token-urile 'we' și 'must' au o influență semnificativă asupra clasificării în dialectul britanic. Acest fapt sugerează că vorbitorii britanici recurg frecvent la persoana întâi plural, subliniind coeziunea colectivă, în timp ce utilizarea cuvântului 'must', cu încărcătură imperativă și retorică puternică, reflectă un stil autoritar al discursului.

Analiza diferențelor în ponderile atribuite articolelor, adverbelor, pronumelor și prepozițiilor – precum 'the', 'that', 'of', 'to', 'in', 'those', 'we', 'now' – arată că aceste clase de cuvinte contribuie majoritar la diferențierea între cele trei dialecte.

Imaginile din dreapta figurii 4.1 ilustrează două exemple de clasificări corecte. Pentru prima, o clasificare foarte încrezătoare, cuvintele determinante sunt 'bureau' și 'house', cuvinte tipic scoțiene conform analizei globale SHAP. În cea de-a doua este ilustrat cum 'british' are o influență pozitivă foarte mare, distrăgând atenția de la potențialele semnale lingvistice.

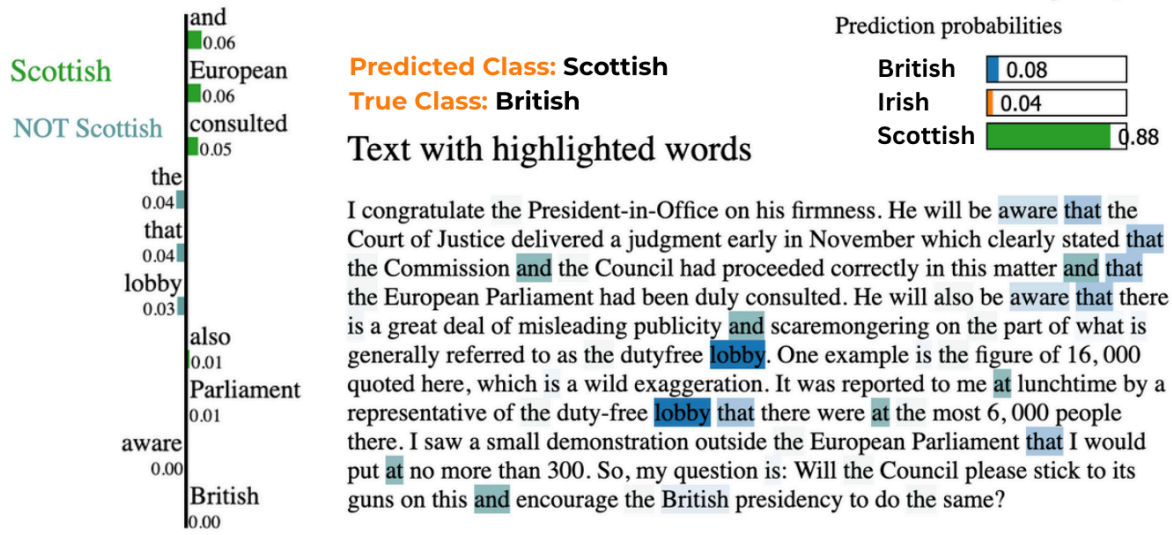

Figura 4.2: Explicarea unei predicții greșit scoțiene a unui text britanic cu Lime.

Totuși, în figura 4.2 este ilustrată explicația oferită de Lime pentru o clasificare eronată a unui text britanic. Deși termenul 'British' este prezent în fragment, influența sa asupra deciziei finale este redusă, iar modelul atribuie în mod greșit textul clasei scoțiene. Acest exemplu sugerează că modelele nu se bazează exclusiv pe cuvinte-cheie evidente, ci iau în considerare o varietate de semnale lingvistice mai subtile în procesul de clasificare.

O ilustrație pentru o predicție corectă a dialectului irlandez o putem urmări în figura 6.5 din anexă. Cele mai relevante cuvinte pentru fragmentul din textul dat sunt 'honourable', 'Ireland', 'the' și 'of'. Pe ultimele trei le putem recunoaște din graficul global SHAP ca având o contribuție majoră în identificarea textelor irlandeze. Această corespondență între analizele locale și cele globale validează și întărește încrederea în interpretabilitatea și corectitudinea modelului.

4.2 Transformers

	Legend: Negative Neutral Positive True Label / Predicted Label / Attribution Label	Word Importance
ModernBERT	British / British (0.84) / British	Mr President, you have already quite properly expressed the sympathy (callous and indiscriminate terrorist unambiguous solidarity.
BERT	British / Irish (0.46) / British	mr president , you have already quite properly expressed the sympathy call ##ous and ind ##is ##cr ##imi ##nate una ##mb ##ig ##uous solidarity

Figura 4.3: Vizualizarea predicțiilor BERT și ModernBERT pentru același text cu CAPTUM.

Pentru a interpreta predicțiile generate de cele două modele de tip transformer, a fost utilizată biblioteca Captum. În Figura 4.3, prima linie ilustrează o clasificare corectă realizată de modelul ModernBERT pentru dialectul britanic, în timp ce a doua linie prezintă o eroare de clasificare a modelului BERT, care etichetează în mod greșit textul ca fiind irlandez. Cuvintele evidențiate cu verde contribuie pozitiv la identificarea clasei britanice, în timp ce cele marcate cu roșu au o influență negativă.

Modelul ModernBERT recunoaște cuvintele 'quite', 'expressed', 'indiscriminate' și 'unambiguous' ca fiind indicativi ai dialectului britanic. În contrast, modelul BERT nu reușește să le asocieze cu această clasă, iar termenii 'unambiguous' și 'indiscriminate' nu sunt recunoscuți de către tokenizer, mai mult, primesc un scor negativ. Aceștia apar ca având o importanță semnificativă pentru clasificarea corectă, conform rezultatelor obținute de ModernBERT.

Capitolul 5

Concluzii

În urma analizei rezultatelor atât calitative cât și cantitative putem formula răspunsuri clare la următoarele întrebări:

1. Există diferențe între dialectele limbii engleze scrise?

Da, cu siguranță exista deosebiri, chiar dacă subtile, în felul în care britanicii, irlandezii și scoțienii își formulează discursul.

2. Dacă da, cum arată acestea?

Distincțiile apar în principal la nivel sintactic – în modul de articulare a propozițiilor și în organizarea cuvintelor. Alegerea lexicală joacă, de asemenea, un rol esențial, în special în utilizarea cuvintelor funcționale precum adverbele, prepozițiile și articolele. Diferențe minore se pot identifica și în utilizarea semnelor de punctuație, deși acestea au o influență mai redusă.

3. Ce tip de clasificatori putem folosi pentru a identifica cel mai eficient aceste diferențe?

Cele mai bune rezultate au fost obținute de către ModernBERT și Support Vector Classifier, care au performat la un nivel aproape identic. SVC se bazează pe separarea geometrică într-un spațiu de trăsături construit statistic, în timp ce ModernBERT beneficiază de un volum masiv de date de antrenament, un tokenizator avansat și o capacitate ridicată de a sesiza nuanțele subtile din limbaj.

4. Ce fel de trăsături (feature-uri) capturează cel mai bine distincțiile dialectale?

Reprezentările Tf-Idf și Bag of Words s-au dovedit eficiente și au oferit performanțe similare pe corpusul analizat. În schimb, embedding-urile de tip sentence sau token-urile CLS, neajustate contextual, nu au reușit să surprindă suficientă informație relevantă, conducând la rezultate mai slabe. Un aspect notabil este că păstrarea cuvintelor de legătură și a punctuației în text contribuie semnificativ la creșterea performanței modelelor.

Observații interesante

Parlamentarilor irlandezi le place cel mai mult să-și afirme identitatea națională. Britanicii folosesc frecvent pronumele 'noi' și au cel mai hotărât discurs, utilizând cuvinte

imperative ca 'trebuie', în timp ce scoțienii sunt cei mai puțin asertivi.

Limitări

Setul de date utilizat este relativ restrâns ca dimensiune și prezintă un stil predominant formal, ceea ce limitează expresivitatea naturală a dialectelor. În plus, tematica discursurilor este destul de uniformă, ceea ce poate favoriza asocierea artificială a claselor cu subiecte recurente, în detrimentul trăsăturilor lingvistice propriu-zise.

Îmbunătățiri propuse

Extinderea și diversificarea setului de date: un corpus mai amplu, care să includă și contexte colocviale sau informale, ar permite o captare mai fidelă a trăsăturilor dialectale autentice.

Integrarea unor trăsături stilometrice — cum ar fi lungimea medie a frazelor sau complexitatea sintactică — ar putea evidenția aspecte fine ale stilului propriu fiecărui dialect.

Analiză bazată pe categorii gramaticale (POS-tagging): Se propune antrenarea modelelor pe subseturi de text filtrate după partea de vorbire — de exemplu, exclusiv substantive, verbe sau adverbe — pentru a investiga care dintre acestea contribuie cel mai mult la diferențierea dialectelor.

Bibliografie

[1] Zaid Al-Jumaili Tarek Bassiouny Ahmad Alanezi Wasiq Khan Dhiya Al-Jumeily Obe Abir Jaafar Hussain, „Classification of Spoken English Accents Using Deep Learning and Speech Analysis", în (2022), URL: https://www.researchgate.net/publication/362705130_Classification_of_Spoken_English_Accents_Using_Deep_Learning_and_Speech_Analysis.

[2] Eric Atwell Junaid Arshad Chien-Ming Lai Lan Nim, „Which English Dominates the World Wide Web, British or American?", în (2007), URL: https://www.researchgate.net/publication/265005347_Which_English_Dominates_the_World_Wide_Web_British_or_American.

[3] Paul Cook Graeme Hirst, „Do Web Corpora from Top-Level Domains Represent National Varieties of English?", în (2012), URL: https://www.cs.toronto.edu/~pcook/CookHirst2012.pdf.

[4] Philipp Koehn Franz J. Och Daniel Marcu, „Europarl (European Parliament Proceedings Parallel Corpus)", în (2003), URL: https://paperswithcode.com/dataset/europarl.

[5] David J. Hand și Keming Yu, „Idiot's Bayes-Not So Stupid After All?", în *International Statistical Review* (2001), URL: https://www.jstor.org/stable/1403452?origin=crossref&seq=1.

[6] Daniel Jurafsky James H. Martin., „Speech and Language Processing. Logistic Regression", în (2025), URL: https://web.stanford.edu/~jurafsky/slp3/5.pdf.

[7] Sebastian Raschka, „Naive Bayes and Text Classification", în (2014), URL: https://arxiv.org/pdf/1410.5329.

[8] Sherrie Wang, „Support Vector Machines", în (2019), URL: https://web.stanford.edu/class/cme250/files/cme250_lecture5.pdf.

[9] scikit-learn developers, *SVC*, https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC, -.

[10] Brownlee Jason, „A Gentle Introduction to k-fold Cross-Validation", în (2023), URL: https://machinelearningmastery.com/k-fold-cross-validation/.

[11] scikit-learn developers, *Feature Selection*, https://scikit-learn.org/stable/modules/feature_selection.html, -.

[12] Thomas Wolf - Huggingface, „The Current Best of Universal Word Embeddings and Sentence Embeddings", în (2018), URL: https://medium.com/huggingface/universal-word-sentence-embeddings-ce48ddc8fc3a.

[13] Momen Ibrahim Ahmed Akram Mohammed Radwan Rana Ayma Mustafa Abd-El-Hameed Nagwa El-Makky Marwan Torki, „Enhancing Authorship Verification using Sentence-Transformers", în *Notebook for PAN at CLEF 2023* (2023), URL: https://ceur-ws.org/Vol-3497/paper-216.pdf.

[14] HuggingFace developers Tom Aarsen, *Sentence Transformer*, https://www.sbert.net/, -.

[15] Aditya Raj, *Understanding the [CLS] Token in BERT: A Comprehensive Guide*, https://aditya007.medium.com/understanding-the-cls-token-in-bert-a-comprehensive-guide-a62b3b94a941, 2024.

[16] Diederik P. Kingma Jimmy Ba, „Adam: A Method for Stochasti Optimization", în (2014), URL: https://arxiv.org/abs/1412.6980v9.

[17] Sergey Ioffe Christian Szegedy, „Batch Normalization: Accelerating Deep Network Training. Reducing Internal Covariate Shift", în (2015), URL: https://arxiv.org/pdf/1502.03167.

[18] Rafael Müller Simon Kornblith Geoffrey Hinton, „When Does Label Smoothing Help?", în (2020), URL: https://arxiv.org/abs/1906.02629v3.

[19] Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", în (2019), URL: https://arxiv.org/pdf/1810.04805.

[20] answer.ai, „Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference", în (2024), URL: https://arxiv.org/pdf/2412.13663.

[21] AnswerAI researchers, *Finally, a Replacement for BERT*, https://huggingface.co/blog/modernbert, 2024.

[22] HuggingFace developers, *Transformers Library*, https://huggingface.co/docs/transformers/en/index, -.

[23] open-source, *SHAP (SHapley Additive exPlanations)*, https://shap.readthedocs.io/en/latest/, 2018.

[24] open-source, *LIME (Local Interpretable Model-agnostic Explanations)*, https://github.com/marcotcr/lime, 2016.

[25] Kokhlikyan N. Miglani V. Martin M. Wang E. Alsallakh B. Reynolds J. Melnikov A. Kliushkina N. Araya C. Yan S. Reblitz-Richardson, *Captum: A unified and generic model interpretability library for PyTorch*, https://doi.org/10.48550/arXiv.2009.07896, 2020.

[26] T. Joachims, „Text categorization with Support Vector Machines: Learning with many relevant features.", în *Machine Learning: ECML-98. ECML 1998* (1998), URL: https://doi.org/10.1007/BFb0026683.

[27] Andrew McCallum și Kamal Nigam, „A Comparison of Event Models for Naive Bayes Text Classification", în (1998), URL: https://aaai.org/papers/041-ws98-05-007/.

[28] Benedikt Szmrecsanyi, „Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry", în (2009), URL: https://www.researchgate.net/publication/220040192_Grammatical_variation_in_British_English_dialects_A_study_in_corpus-based_dialectometry.

[29] David John Britain, „Grammatical variation in the contemporary spoken English of England", în (2010), URL: https://www.researchgate.net/publication/260087978_Grammatical_variation_in_the_contemporary_spoken_English_of_England.

Capitolul 6

Anexă

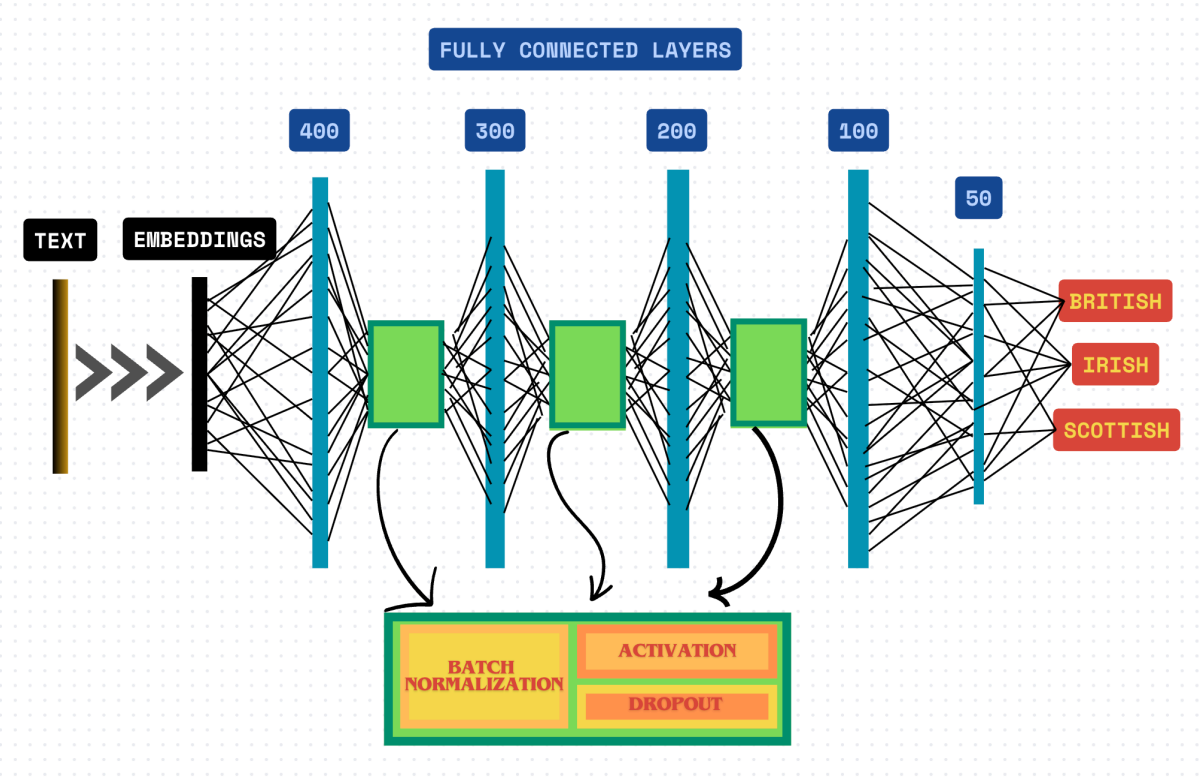

Figura 6.1: Ilustrație pentru arhitectura rețelei neuronale folosite.

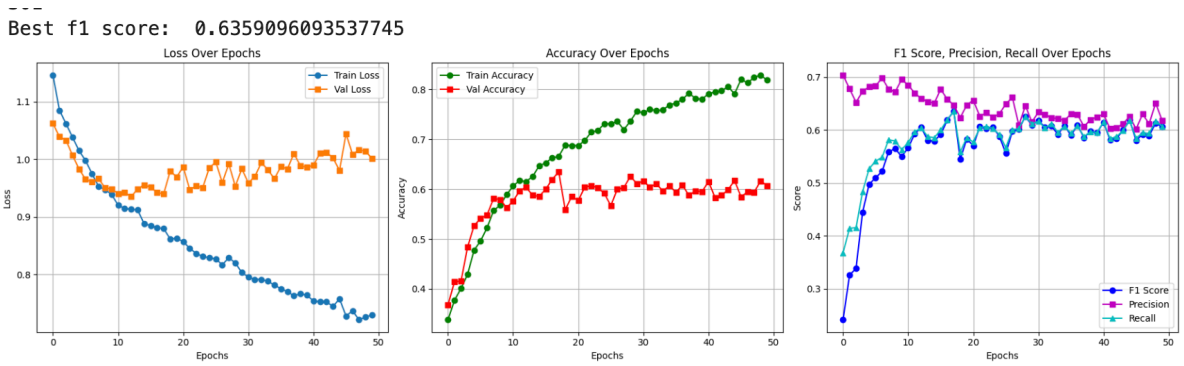

Figura 6.2: Curbe de antrenare a rețelei neuronale pentru token-ul CLS din BERT.

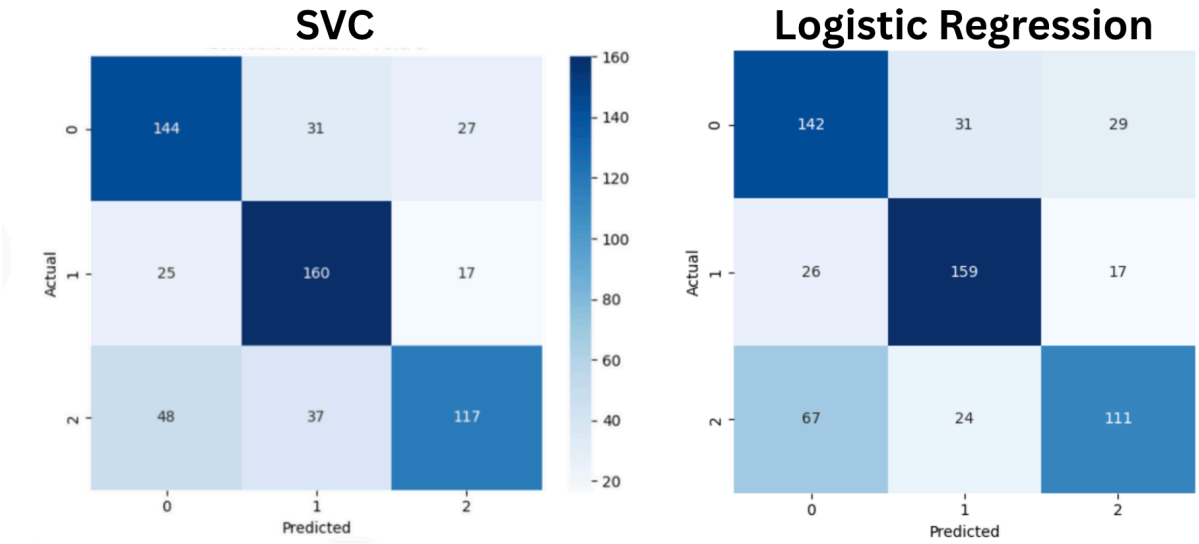

Figura 6.3: Matrice de confuzie pentru Logistic Regression și SVC. Clasele sunt după cum urmează: 0 - British, 1 - Irish, 3 - Scottish.

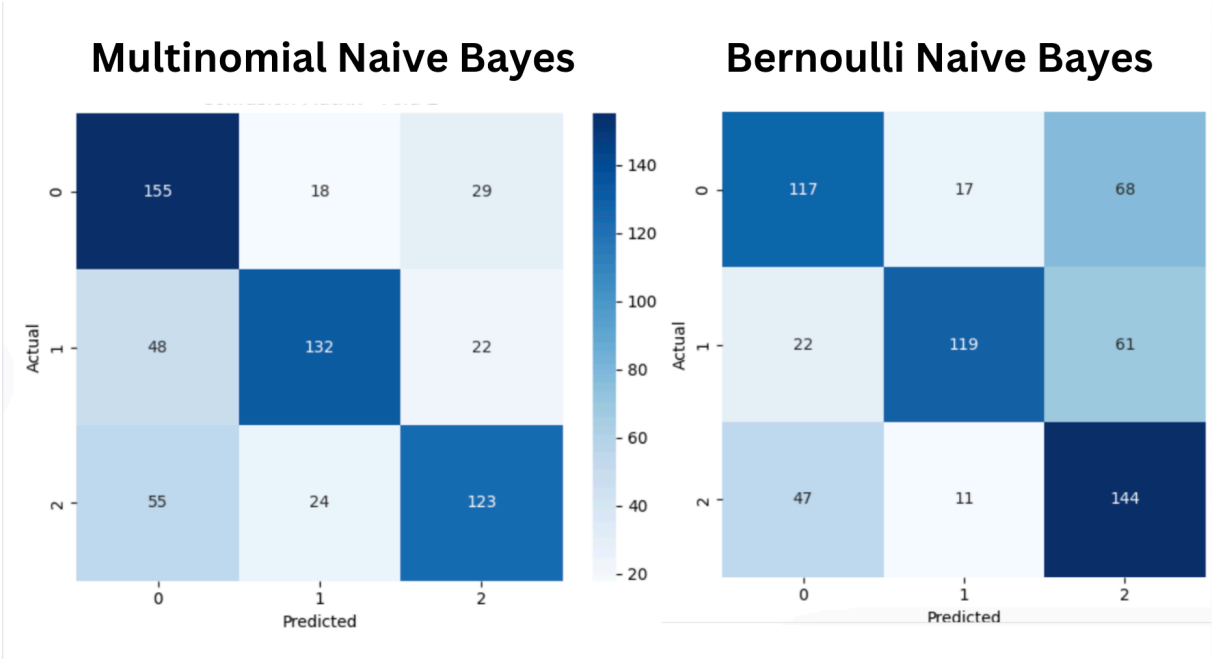

Figura 6.4: Matrice de confuzie pentru Logistic modelele Naive Bayes. Clasele sunt după cum urmează: 0 - British, 1 - Irish, 3 - Scottish.

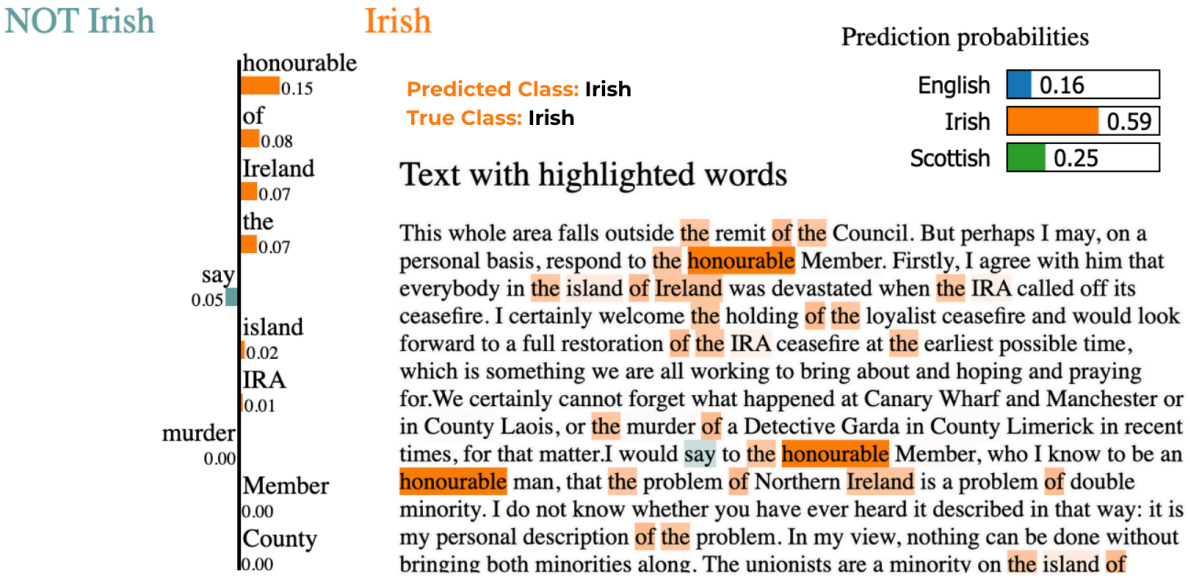

Figura 6.5: Explicarea unei predicții irlandeze corecte cu Lime.