Datarea textelor din Limba Romana

Mihailescu Teodor teodor.mihailescu@s.unibuc.ro

Ionescu Radu-Constantin radu-constantin.ionescu@s.unibuc.ro

Abstract

Scopul acestei lucrari este dezvoltarea si analizarea unui model de Inteligenta Artificiala pentru datarea textelor din limba romana, din secolul al XIII-lea pana in ziua de azi.

1 Introduction

Problema data are o importanta practica mai mult decat teoretica. Ca o consecinta directa a contextului istoric al poporului si limbii romane, izvoarele istorice se imputineaza exponential invers proportional cu timpul, ceea ce se traduce in date de antrenare nebalansate si insuficiente pentru anumite secole. Metodologia istoricilor de a data un text se bazeaza pe identificarea anumitor ancore temporale in texte (anumite denumiri pentru elemente geografice, personalitati istorice, evenimente pentru care se cunoaste spatiul temporal). Pe de alta parte, abordarea lingvistilor se axeaza pe corelarea anumitor elemente de limbaj (vocabular, gramatica, ortografie) cu perioada lor de utilizare pe parcursul evolutiei limbii romane. In ajutorul celor din urma aceasta lucrare isi propune sa vina in ajutor, incercand automatizarea acestui proces.

2 Related Work

Tema in cauza a fost tratata excelent in urmatoarele:

- Temporal classification for historical Romanian texts: Alina Maria Ciobanu, Liviu P. Dinu, Octavia-Maria Sulea, Anca Dinu, Vlad Niculae
- Temporal Text Classification for Romanian Novels Set in the Past: Alina Maria Ciobanu, Liviu P. Dinu, Octavia-Maria Sulea, Anca Dinu, Vlad Niculae

3 Method

La prima vedere, problema pare una de clasificare, insa urmatorul detaliu in legatura cu clasele predictorilor indica o noua dimeniune a problemei: intre secolele reprezentate ca numere exista o relatie de ordine. Astfel, chiar daca ideal ar fi ca modelul sa indice perfect clasele reale ale textelor, o predictie mai arpopiata de realitate este preferabila in locul uneia la distanta de mai multe secole. Intr-un fel, problema pare acum una de regresie si poate fi privita si asa, insa din cosniderente de vizualizare aceasta lucrare isi propune sa trateze problema ca una de clasificare.

Astfel, data fiind natura problemei, in vederea evaluarii performantei solutiilor propuse, s-a folosit o metrica peronalizata de evaluare, numita in continuare ***Acuratete Ponderata (Weighted Accuracy - wacc[2])***

$$wacc(f, X, y, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \gamma^{-|f(X^T)_i - y_i|}$$

- $f : \mathbb{R}^d \to \mathbb{C}$ unde $\mathbb{C} \subseteq \mathbb{N}$ este multimea claselor, iar d reprezintă dimensiunea spațiului de vectorizare.
- $X \in \mathbb{R}^{n \times d}$, unde $n \in \mathbb{N}$ (numarul de esantioane).
- $y \in \mathbb{N}^n$, reprezinta etichetele datelor
- $\gamma \in \mathbb{R}, \gamma > 1$, reprezinta factorul de penalizare

Se foloseste aceasta metrica deoarece are urmatoarele proprietati:

$$\lim_{\gamma \to \infty} wacc(f, X, y, \gamma) = acc(f, X, y)$$

$$wacc(f, X, y, \gamma_2) > wacc(f, X, y, \gamma_1)$$

$$pentru\ \gamma_1, \gamma_2 \in R, \gamma_1, \gamma_2 > 1, \gamma_1 > \gamma_2$$

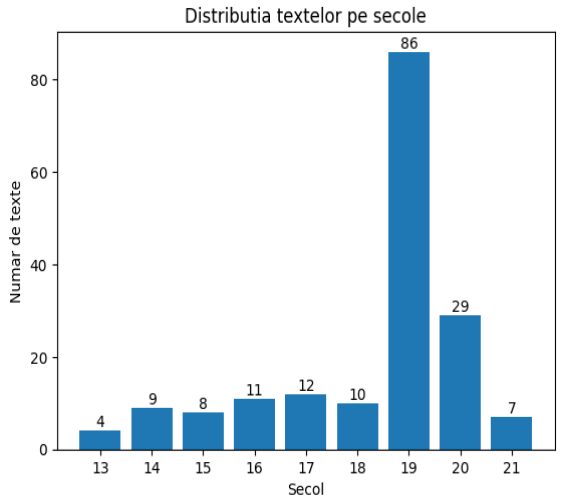

Figure 1: Distributia Datelor

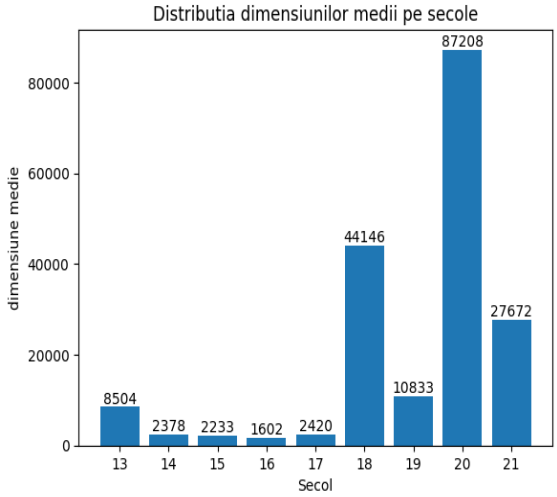

Figure 2: Distributia Dimensiunii Textelor

3.1 Dataset

In vederea antrenarii modelelor, s-a folosit un set de date inspirat partial din lucrarea de referinta de aici, plus alte traduceri ale anumitor documente istorice din texte istorice, de dinainte epocii literare. Dupa cum s-a mentionat si in introducere, setul de date este nebalansat, urmand distrubutia din Figura 1. dupa numarul de exemple de antrenare si distributia din Figura 2 dupa dimeniunea medie a textelor.

Din aceasta distributie, se observa ca o anumita clasa domina setul de date, instantele sale constituind ***48.8%*** din totalul datelor, acesta devenind ***baseline-ul***.

Din setul de date original, se distribuie 80% din date spre setul de antrenare si 20% spre setul de validare.

3.2 Preprocessing

In vederea transformarii textelor in informatii numerice, se foloseste vectorizarea bazata pe frecevnta inversa relativa la celelate documente: ***Vectorizarea TFIDF***

3.3 Models

Pentru clasificarea textelor s-a experimentat cu 5 tipuri de modele de invatare automata, fara retele neuronale. Pentru fiecare model s-a realizat tunarea hiperparametrilor pentru a obtine rezultate cat mai bune.

Model	wacc
SVC[4]	0.85
KNN[3]	0.74
Decission Tree[6]	0.69
Naive Bayes (Multinomial)[1]	0.69
Naive Bayes (Gaussian)[5]	0.64

4 Conclusion

Acest proiect a reprezintat atat o oprtunitate de invatare si dezvoltare pentru noi, cat si o speranta ca munca noastra s-ar putea sa ajute un grup de oameni in desfasurarea activitatii lor profesionale.

5 Future Work

In primul rand, modele ar putea fi rafinate prin colectarea mai multor date si tunarea hiperparametrilor pe noul set de date. Ar fi interesant ca acest clasificator sa faca parte dintr-un GAN cu ajutorul caruia s-ar putea genera texte din limba romana a secolelor trecute, eventual si din perioadele mai arhaice pentru care nu exista izvoare istorice, astfel putand sa fie facut un fel de "reverse engineering" pe limba romana actuala si din secolele trecute pentru a afla cum arata mai exact limba romana din secolele 8-12 de exemplu.

6 Ethical Statement

In dezvoltarea si analiza modelelor s-a constatat, doar la nivel observational, un bias in perceptia asupra datelor: textele de secol 18 erau predominant texte bisericesti, cele de secol 19 literatura, in special poiezie, cele de secol 20 documente administrative, declaratii oficiale. Desi s-ar fi putut rezolva problema clasificarii la nivel de secol cu o generalitate destul de buna doar antrenand un clasificator care sa decida daca textul vorbeste sau nu despre subiectul principal comun al secolului,

acest lucru a fost evitat. De ce? Deoarece nu se poate porni la drum cu prejudecata ca toata limba romana scrisa in secolul al 18-lea este despre psaltiri; poate ca intr-o zi cineva va descoperi o colectie nepublicata, nedatata de poezii ale unui individ ce a trait in secolul respectiv, poezii al caror subiect nu ar fi neaparat despre divinitate, si astfel ar fi etichetate gresit din simpla ignoranta.

Astfel aceasta problema de etica a fost tratata aplicand acelasi tip de preprocesare tuturor datelor si folosind un clasificator universal pentru toate secolele, acesta fiind mult mai putin predispus la identificarea uunui tipar similar prejudecatilor, in schimb axandu-se pe caracteristicle mai generale ale textelor.

Limitations

In mod evident, modelele dezvoltate au fost mulate pe setul de date al limbii romane. Nu exista garantia functionarii acestui clasificator pe texte din alte limbi cu scopul datarii corecte a textelor.

Acknowledgements

Multumim in mod special domnului doctor conferentiar universitar Liviu P. Dinu pentru informatiile acordate pe parcursul acestui curs si inspiratia oferita din lucrarile la care dumnealui a contribuit!

References

[1] Muhammad Abbas, K Ali Memon, A Aleem Jamali, Saleemullah Memon, and Anees Ahmed. “Multinomial Naive Bayes classification model for sentiment analysis”. In: *IJCSNS Int. J. Comput. Sci. Netw. Secur* 19.3 (2019), p. 62.

[2] Ivan Evdokimov, Michael Kampouridis, and Tasos Papastylianou. “Application Of Machine Learning Algorithms to Free Cash Flows Growth Rate Estimation”. In: *Procedia Computer Science* 222 (2023), pp. 529–538.

[3] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. “KNN model-based approach in classification”. In: *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*. Springer. 2003, pp. 986–996.

[4] Pubudu L Indrasiri, Malka N Halgamuge, and Azeem Mohammad. “Robust ensemble machine learning model for filtering phishing URLs: Expandable random gradient stacked voting classifier (ERG-SVC)”. In: *Ieee Access* 9 (2021), pp. 150142–150161.

[5] Eguturi Manjith Kumar Reddy, Akash Gurrala, Vasireddy Bindu Hasitha, and Korupalli V Rajesh Kumar. “Introduction to Naive Bayes and a review on its subtypes with applications”. In: *Bayesian reasoning and gaussian processes for machine learning applications* (2022), pp. 1–14.

[6] Shan Suthaharan and Shan Suthaharan. “Decision tree learning”. In: *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning* (2016), pp. 237–269.