

**MINISTERUL EDUCAȚIEI ȘI CERCETĂRII AL REPUBLICII MOLDOVA**

**Universitatea Tehnică a Moldovei**

**Facultatea Calculatoare, Informatică și Microelectronică**

**Departamentul Ingineria Software și Automatică**

**Admis la susținere**

**Șef departament:**

**Ion Fiodorov, conf. univ., dr.**

” ” \_\_\_\_\_ 2022

## **Augmentarea structurală și semantică a textului**

**Teză de master**

**Student: Prijilevschi Dumitru, TI-201M**

**Conducător: Chirev Pavel, lector univ., dr.**

**Consultant: Cojocaru Svetlana, lector univ. mag.**

**Chișinău, 2022**

## Rezumat

Aceasta lucrare reprezintă o analiză a metodelor de augmentare semantică și structurală a textului. Analiza acestor metode se caracterizează prin cercetarea teoretică, implementare și analiă a rezultatelor asupra testelor. Forma de reprezentare a rezultatelor este realizata într-o pagină web în care sa implementat funcționalul de baza pentru a vedea comportamentul metodelor de augmentare structurală și semantică realizate. Metodele au fost divizate în doua categorii deoarece ideile metodelor au scopuri de a schimba structura textului sau semantica lui. Rezultatele acestor metode au fost comparate între ele la niver de mecanism și la niver de categorie (semantic și structural).

Această lucrare este divizată în 4 capitole care arată un proces de cercetare a unei probleme. În capitolul 1 este argumentată tema de cercetare prin expunerea statisticii ocuparii în domeniu realizata de organizații specializate, care arată sporirea interesului față de domeniu, fapt ce motivează la realizarea lucrarii. În capitolul 2 este realizată analiza teoretică a temei și este capitolul cel mai important. Aici se analizează detaliat începînd cu domeniul aferent și terminînd cu analiza fiecărei metode în parte. În capitolul 3 sa implementat metodele de augmentare semantică și structurală și sa descris procesul de realizare și instrumentele folosite. În capitolul 4 sa realizat un prototip care implementeaza toate metodele realizate cu o interfață comodă de interacțiune sub forma unei aplicații web. Concluziile acestei lucrări descriu rezultatele optținute și recomandări față de metode.

## **Abstract**

This paper is an analysis of the methods of semantic and structural augmentation of the text. The analysis of these methods is characterized by theoretical research, implementation and analysis of test results. The form of representation of the results is made in a web page in which the basic functional has been implemented to see the behavior of the structural and semantic augmentation methods performed. Methods have been divided into two categories because the ideas of the methods are intended to change the structure of the text or its semantics. The results of these methods were compared with each other at the mechanism level and at the category level (semantic and structural).

This paper is divided into 4 chapters that show a process of researching a problem. Chapter 1 argues the research topic by exposing the employment statistics in the field by specialized organizations, which shows increased interest in the field, which motivates the work. Chapter 2 is the theoretical analysis of the topic and is the most important chapter. Here we analyze in detail starting with the related field and ending with the analysis of each method. Chapter 3 implemented the methods of semantic and structural augmentation and described the implementation process and tools used. In Chapter 4 we made a prototype that implements all the methods made with a convenient interaction interface in the form of a web application. The conclusions of this paper describe the results obtained and recommendations for methods.

## CUPRINS

INTRODUCERE.....	5
1. ARGUMENTAREA TEMEI DE CERCETARE .....	6
1.1 INVESTIȚIILE GLOBALE ÎN DOMENIUL AI ȘI NLP.....	6
1.2 ACTIVITATEA DE CERCETARE ȘI DEZVOLTARE ÎN DOMENIUL AI ȘI NLP .....	7
2. ANALIZA DOMENIULUI .....	9
2.1 PROCESAREA LIMBAJULUI NATURAL .....	10
2.2 AUGMENTAREA TEXTULUI .....	12
2.3 ANALIZA METODELOR DE AUGMETARE STRUCTURALĂ.....	12
2.4 ANALIZA METODELOR DE AUGMENTARE SEMANTICĂ.....	13
3. IMPLEMENTAREA METODELOR ANALIZATE.....	16
3.1 IMPLEMENTAREA METODELOR DE AUGMENTARE .....	18
4. IMPLEMENTAREA PROTOTIPULUI DE AUGMENTARE A TEXTULUI .....	21
4.1 DOCUMENTAREA APLICAȚIEI .....	21
BIBLIOGRAFIE.....	24

## INTRODUCERE

În procesere de luare a deciziilor business se opservă tendința de a folosi analiza multor date. Analiza datelor reprezintă un proces în care informația acumulată este curățită sau transformată, după care este reprezentată într-o formă specifică (ex: grafică, schematică) pentru a fi mai comod pentru om de a înțelege aceste date și a face o concluzie. Inteligența artificială (AI) este o tehnologie care are deja un impact asupra modului în care utilizatorii interacționează cu internetul și sunt afectați de acesta. În viitorul apropiat, impactul său va continua să crească. AI are potențialul de a schimba foarte mult modul în care oamenii interacționează, nu numai cu lumea digitală, ci și între ei, prin munca lor și prin alte instituții socio-economice în bine sau în rău. Inteligența artificială (AI) este o tehnologie care avansează rapid, făcută posibilă de internet, care poate avea în curând un impact semnificativ asupra vieții noastre de zi cu zi. AI se referă în mod tradițional la o creație artificială a inteligenței umane care poate învăța, raționa, planifica, percepe sau procesa limbajul natural<sup>1</sup>. Aceste trăsături permit AI să aducă oportunități socio-economice imense, punând în același timp provocări etice și socio-economice. Inteligența artificială (AI) a primit o atenție sporită în ultimii ani. Inovația, posibilă prin intermediul internetului, a adus AI mai aproape de viața noastră de zi cu zi. Aceste progrese, alături de interesul pentru potențialele impacturi socio-economice și etice ale tehnologiei, aduc AI în prim-planul multor debateri contemporane. Investițiile din industrie în AI cresc rapid, iar guvernele încearcă să înțeleagă ce ar putea însemna tehnologia pentru cetățenii lor. Procesarea limbajului natural (NLP) a câștigat recent multă atenție pentru reprezentarea și analiza computerizată a limbajului uman. Și-a răspândit aplicațiile în diverse domenii, cum ar fi traducerea automată, detectarea e-mail-urilor, extragerea informațiilor, rezumarea, medicale și răspunsuri la întrebări etc. prin prezentarea istoriei și evoluției NLP, stare de artă prezentând diferitele aplicații ale NLP și tendințele și provocările actuale.

## BIBLIOGRAFIE

- [1] D., Zhang, S., Mishra, E., Brynjolfsson, J., Etchemendy, D., Ganguli, B., Grosz, T., Lyons, J., Manyika, J., Carlos Niebles, M., Sellitto, Y., Shoham, J., Clark, și R., Perrault, “The AI Index 2021 Annual Report”, Comitetul AI Index Steering, Institutul Human-Centered AI, Universitatea Stanford, Stanford, CA, USA, 2021 mar., pp 14-100.
- [2] D. Khurana, A. Koli, K. Khatter și S. Singh, Natural Language Processing: State of The Art, Current Trends and Challenges, Departamentul Computer Science and Engineering, Universitatea Internațională Manav Rachna, India.
- [3] C. Shorten, T. M. Khoshgoftaar și B. Furht, “Text Data Augmentation for Deep Learning”, Journal of Bid Data, pp 4-6, DOI: [10.1186/s40537-021-00492-0](https://doi.org/10.1186/s40537-021-00492-0)
- [4] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Hugging Face, Brooklyn, SUA
- [5] S. Y. Feng, V. Gangal, D. Kang, T. Mitamura, E. Hovy, “GenAug: Data Augmentation for Finetuning Text Generators”, Universitatea Carnegie Mellon și Universitatea din California, Berkeley, 2020