MINISTERUL EDUCAŢIEI ŞI CERCETĂRII AL REPUBLICII MOLDOVA
Universitatea Tehnică a Moldovei
Facultatea Calculatoare Informatică şi Microelectronică
Departamentul Ingineria Software și Automatică

Admis la susţinere
Şef departament:
FIODOROV Ion dr., conf.univ.

„___” _____ 2025

PROCESE DE PRELUCRARE A LIMBAJULUI NATURAL PENTRU ANALIZA DATELOR TEXTUALE DIN REȚELELE SOCIALE

Proiect de master

Student: _____ **Popa Ana, TI-231M**

Coordonator: _____ **Cojocaru Svetlana, asis. univ.**

Consultant: _____ **Cojocaru Svetlana, asis. univ.**

Chişinău, 2025

REZUMAT

Această teză abordează procesarea limbajului natural (PLN) pentru analiza datelor textuale generate de rețelele sociale, cu scopul de a dezvolta metode automatizate de colectare, preprocesare și interpretare a datelor. Lucrarea este structurată în mai multe capitole, fiecare având un rol specific în fundamentarea teoretică și aplicativă a subiectului.

Introducerea oferă contextul și relevanța studiului, subliniind importanța utilizării PLN pentru gestionarea volumelor mari de date textuale și transformarea acestora în informații valoroase. Se definesc scopul și obiectivele tezei, evidențiind nevoia de soluții automatizate pentru analiza textului din rețelele sociale.

Capitolul 1 analizează procesele fundamentale de prelucrare a limbajului natural. Sunt descrise contextul și importanța PLN, tehnicile și metodele utilizate, precum și aplicațiile acestora în diverse domenii. De asemenea, se discută specificul textului din rețelele sociale și provocările pe care le prezintă în analiza automată.

Capitolul 2 se concentrează pe metodele de colectare și pregătire a datelor din rețelele sociale. Se prezintă sursele de date, clasificarea platformelor sociale și tipurile de date disponibile, inclusiv text, metadate și conținut multimedia. Se discută metodele de preprocesare, precum eliminarea zgomotului, tokenizarea și normalizarea, alături de provocările etice și tehnice asociate acestora.

Concluziile sintetizează contribuțiile și rezultatele obținute, subliniind importanța tehnologiilor de PLN pentru analiza textului din rețelele sociale. Sunt identificate direcții viitoare de cercetare și provocările care trebuie abordate pentru îmbunătățirea aplicațiilor existente.

Lucrarea aduce o contribuție în domeniul procesării limbajului natural, oferind un cadru teoretic și practic pentru analiza datelor textuale din rețelele sociale, cu aplicații în marketing, cercetare socială și alte domenii relevante.

ABSTRACT

This thesis explores natural language processing (NLP) for analyzing textual data generated from social networks, aiming to develop automated methods for data collection, preprocessing, and interpretation. The thesis is structured into several chapters, each contributing to the theoretical and practical understanding of the subject.

The introduction provides the context and relevance of the study, highlighting the importance of using NLP to manage large volumes of textual data and transform them into valuable insights. The objectives and purpose of the thesis are defined, emphasizing the need for automated solutions for analyzing social media text.

Chapter 1 examines the fundamental processes of natural language processing. It describes the context and significance of NLP, the techniques and methods used, as well as its applications across various fields. Additionally, it discusses the specific characteristics of social media text and the challenges it presents for automated analysis.

Chapter 2 focuses on methods for collecting and preparing data from social networks. It presents data sources, the classification of social platforms, and the types of data available, including text, metadata, and multimedia content. The chapter covers preprocessing methods such as noise reduction, tokenization, and normalization, alongside ethical and technical challenges associated with these processes.

The conclusions summarize the contributions and results, emphasizing the significance of NLP technologies for analyzing social media text. Future research directions and challenges to be addressed for improving current applications are also identified.

This thesis makes a contribution to the field of natural language processing, offering both theoretical and practical frameworks for analyzing textual data from social networks, with applications in marketing, social research, and other relevant domains.

CUPRINS

ABREVIERI

NLP - Natural Language Processing.

PLN - Prelucrarea Limbajului Natural.

API - Application Programming Interface.

GPU - Graphics Processing Unit.

BERT - Bidirectional Encoder Representations from Transformers.

GPT - Generative Pre-trained Transformer.

RNN - Recurrent Neural Network.

LSTM - Long Short-Term Memory.

CNN - Convolutional Neural Network.

LDA - Latent Dirichlet Allocation.

SVM - Support Vector Machine.

NLTK - Natural Language Toolkit.

TF-IDF - Term Frequency-Inverse Document Frequency.

OCR - Optical Character Recognition.

INTRODUCERE

Într-o eră dominată de digitalizare, rețelele sociale joacă un rol central în comunicarea și interacțiunea umană, generând zilnic un volum masiv de date textuale. Aceste platforme oferă o oglindă a gândurilor, emoțiilor și opiniilor utilizatorilor, constituind astfel o sursă valoroasă pentru analizarea comportamentelor și tendințelor sociale. Totodată, creșterea exponențială a acestor date impune utilizarea unor tehnologii avansate pentru procesarea și interpretarea lor, astfel încât să poată fi transformate în informații utile.

Prelucrarea Limbajului Natural (PLN), o ramură a inteligenței artificiale, reprezintă soluția la aceste provocări, oferind instrumente și tehnici care permit analiza automată a datelor textuale. În combinație cu algoritmi de învățare automată și arhitecturi de rețele neuronale, PLN facilitează extragerea de cunoștințe din texte nestructurate, îmbunătățind înțelegerea interacțiunilor umane. În acest context, rețelele sociale devin nu doar un mediu de comunicare, ci și un spațiu de cercetare și inovație.

Această teză de master își propune să exploreze procesele și metodele de prelucrare a limbajului natural aplicate în analiza datelor din rețelele sociale, având ca obiectiv principal dezvoltarea unor abordări automatizate pentru colectarea, preprocesarea și interpretarea datelor textuale. Prin utilizarea tehnologiilor moderne, cum ar fi modelele de limbaj avansate (e.g., BERT, GPT), se urmărește obținerea unor perspective detaliate asupra opiniilor utilizatorilor, analiza sentimentelor și identificarea tendințelor emergente.

Relevanța acestui subiect este dublată de complexitatea datelor textuale generate de rețelele sociale, caracterizate de limbaj informal, diversitate lingvistică și dinamica rapidă a conținutului. Prin abordarea provocărilor specifice acestui domeniu, cum ar fi gestionarea volumelor mari de date, diversitatea culturală sau respectarea confidențialității utilizatorilor, lucrarea contribuie la îmbunătățirea tehnologiilor existente și la dezvoltarea unor soluții adaptabile pentru analiza textului în medii digitale.

Integrarea metodelor avansate de prelucrare a limbajului natural în analiza datelor din rețelele sociale deschide noi orizonturi în cercetare și aplicare, oferind organizațiilor și cercetătorilor posibilitatea de a valorifica datele textuale în moduri inovatoare. Teza se concentrează pe investigarea tehnologiilor, metodelor și instrumentelor specifice acestui domeniu, contribuind astfel la progresul tehnologic și la înțelegerea complexității interacțiunilor sociale din mediul online.

BIBLIOGRAFIE

[1] Boyd, Danah M., and Nicole B. Ellison. "Social Network Sites: Definition, History, and Scholarship." *Journal of Computer-Mediated Communication*, vol. 13, no. 1, 2007, pp. 210–30, https://doi.org/10.1111/j.1083-6101.2007.00393.x.

[2] Cambria, Erik, and Bebo White. "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]." *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, May 2014, pp. 48–57, https://doi.org/10.1109/MCI.2014.2307227.

[3] Chomsky, Noam. *Language and Mind*. 3. ed. Reprinted, Cambridge University Press, 2007.

[4] Ferragina, Paolo, and Ugo Scaiella. "Fast and Accurate Annotation of Short Texts with Wikipedia Pages." *IEEE Software*, vol. 29, no. 1, Jan. 2012, pp. 70–75. https://doi.org/10.1109/MS.2011.122.

[5] Golder, Scott A., and Michael W. Macy. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures." *Science*, vol. 333, no. 6051, Sept. 2011, pp. 1878–81, https://doi.org/10.1126/science.1202775.

[6] Hirschberg, Julia, and Christopher D. Manning. "Advances in Natural Language Processing." *Science*, vol. 349, no. 6245, July 2015, pp. 261–66, https://doi.org/10.1126/science.aaa8685.

[7] *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781, arXiv, 7 Sept. 2013, https://doi.org/10.48550/arXiv.1301.3781.

[8] Mikolov, Tomas, et al. *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781, arXiv, 7 Sept. 2013, https://doi.org/10.48550/arXiv.1301.3781.

[9] *Introduction to Information Retrieval*. Reprint., Cambridge University Press, 2017.

[10] Zafarani, Reza. *Social Media Mining*. 1st ed, Cambridge University Press, 2014.

[11] Kaplan, Andreas M., and Michael Haenlein. "Users of the World, Unite! The Challenges and Opportunities of Social Media." *Business Horizons*, vol. 53, no. 1, Jan. 2010, pp. 59–68, https://doi.org/10.1016/j.bushor.2009.09.003.

[12] Jurafsky, Dan, and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2. ed. [Nachdr.], Prentice Hall, 2009.