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.

„___” _____ 2023

Managementul procesului de extragere dinamica a datelor pentru analiza sentimentelor

Teză de master

Student: _____ **Dimitriev Alexandru, TIA-211M**

Coordonator: _____ **Nistor Grozavu, Conf.univ., dr.**

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

Chişinău, 2023

REZUMAT

Această teză de master propune o nouă abordare a analizei sentimentelor bazată pe data mining dinamic. Scopul studiului este de a explora utilizarea tehnicilor de extragere a datelor în timp real pentru a îmbunătăți acuratețea și actualitatea analizei sentimentelor. Metodologia de cercetare implică o revizuire a literaturii de specialitate a tehnicilor existente de analiză a sentimentelor și dezvoltarea unui model dinamic de data mining folosind instrumente de web scraping.

Modelul a fost testat pe un set de date de tweet-uri politice colectate în timpul campaniei pentru alegerile prezidențiale din Franța. Rezultatele arată că utilizarea exploatării dinamice a datelor poate îmbunătăți în mod semnificativ acuratețea și actualitatea analizei sentimentelor, oferind o perspectivă valoroasă asupra preferințelor publice și a opiniilor politice. Analiza sentimentelor poate fi folosită în campaniile electorale pentru a identifica teme importante și pentru a obține o mai bună înțelegere a preferințelor electoratului.

Rezultatele studiului arată că extragerea dinamică a datelor poate îmbunătăți în mod semnificativ acuratețea și actualitatea analizei sentimentelor, oferind informații valoroase pentru companii și organizații. Abordarea propusă poate fi aplicată în diverse domenii, inclusiv marketing, servicii pentru clienți și monitorizarea opiniei publice. Studiul subliniază, de asemenea, importanța colectării și analizei continue a datelor.

În general, acest studiu oferă o nouă perspectivă asupra analizei sentimentelor și demonstrează potențialul exploatării dinamice a datelor ca instrument puternic pentru îmbunătățirea acurateței și actualității analizei sentimentelor. Sunt necesare cercetări suplimentare pentru a explora scalabilitatea și robustețea abordării propuse și pentru a identifica potențialele limitări și provocări.

ABSTRACT

This master's thesis proposes a new approach to sentiment analysis based on dynamic data mining. The aim of the study is to explore the use of real-time data mining techniques to improve the accuracy and timeliness of sentiment analysis. The research methodology involves a literature review of existing sentiment analysis techniques and the development of a dynamic data mining model using web scraping tools.

The model was tested on a dataset of political tweets collected during the French presidential election campaign. The results show that using dynamic data mining can significantly improve the accuracy and timeliness of sentiment analysis, providing valuable insight into public preferences and political opinions. Sentiment analysis can be used in election campaigns to identify important themes and gain a better understanding of the electorate's preferences.

The study results show that dynamic data mining can significantly improve the accuracy and timeliness of sentiment analysis, providing valuable insights for companies and organizations. The proposed approach can be applied in various fields, including marketing, customer service and public opinion monitoring. The study also highlights the importance of continuous data collection and analysis.

Overall, this study provides a new insight into sentiment analysis and demonstrates the potential of dynamic data mining as a powerful tool for improving the accuracy and timeliness of sentiment analysis. Further research is needed to explore the scalability and robustness of the proposed approach and to identify potential limitations and challenges.

CUPRINS

LISTA ABREVIERILOR

AI - Artificial Intelligence (Inteligență Artificială)

NLP - Procesarea limbajului natural

ML - Machine Learning (Invatare automata)

SA - Sentiment Analysis (Analiza sentimentelor)

SVM - Support Vector Machine (Mașină cu vectori de suport)

PCA - Principal Component Analysis (Analiza Componentelor Principale)

HMM - Hidden Markov Model (Modelul Markov Ascuns)

LDA - Linear Discriminant Analysis (Analiza Discriminantului Liniar)

QDA - Quadratic Discriminant Analysis (Analiza Discriminantului Pătratic)

BERT - Bidirectional Encoder Representations from Transformers (Reprezentări ale codificatorului bidirecțional din transformatori)

MLM - Masked Language Modeling

SGD - Stochastic Gradient Descent (Algoritm de optimizare)

NMF - Non-negative Matrix Factorization (tehnica de factorizare a matricilor)

TF-IDF (Term Frequency-Inverse Document Frequency)

INTRODUCERE

Această lucrare de cercetare abordează tema detectării de sentimente, o problemă importantă în domeniul analizei datelor și al înțelegerii semnificației mesajelor transmise prin intermediul platformelor online. Scopul acestei lucrări este de a explora diferite abordări și tehnici pentru a detecta sentimentele din mesajele textuale și de a evalua eficiența lor în ceea ce privește acuratețea și performanța.

În contextul creșterii rapide a cantității de date generate de utilizatori pe platformele online, detectarea de sentimente devine o sarcină din ce în ce mai importantă și mai complexă. Această lucrare abordează provocarea analizei unui volum mare de date și a dezvoltat o metodă de detectare a sentimentelor bazată pe un model de învățare automată.

În această lucrare de cercetare, am analizat diverse abordări și tehnici de detectare a sentimentelor, precum și diferite seturi de date, inclusiv mesaje textuale și recenzii ale clienților. Am folosit tehnici de preprocesare a datelor, precum tokenizarea și eliminarea stop-words, și am implementat diferite modele de învățare automată, inclusiv modele de clasificare a sentimentelor și modele de rețele neuronale.

Rezultatele obținute au demonstrat că abordarea noastră bazată pe un model de învățare automată a obținut o acuratețe remarcabilă în detectarea de sentimente din mesajele textuale, cu o performanță mai bună decât alte metode de analiză a sentimentelor utilizate în prezent. Concluziile acestui studiu arată că tehnologiile bazate pe învățarea automată au un potențial imens în ceea ce privește detectarea de sentimente și că abordarea noastră poate fi utilizată cu succes în diverse domenii.

BIBLIOGRAFIE

1. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022, 2003.
2. Peter Cheeseman, James Kelly, Matthew Self, John Stutz, Will Taylor, and Don reeman. Autoclass: A bayesian classification system. In Fifth International Workshop on Machine learning, p. 54–64, 1988.
3. Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. Journal of the American society for information science, 41(6):391, 1990.
4. Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pages 1188–1196, 2014.
5. Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In hlt-Naacl, volume 13, p. 746–751, 2013.
6. R.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics., 20:53–65, 1987.
7. N. Grozavu, N. Rogovschi, and L. Lazhar. Spectral clustering trought topological learning for large datasets. In Neural Information Processing - 23rd International Conference, ICONIP, Proceedings, Part III, p. 119–128, 2016.
8. Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In Proceedings of the 2005 SIAM International Conference on Data Mining, p 606–610. SIAM, 2005.
9. Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In Advances in neural information processing systems, p. 3294–3302, 2015.
10. Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In Advances in neural information processing systems, p. 556–562,2001.
11. "Twitter Sentiment Analysis: The Good the Bad and the OMG!" de Saif, H., Fernandez, M., He, Y. și Alani, H. - publicată în Proceedings of the International AAAI Conference on Weblogs and Social Media, 2013.
12. "A Sentiment Analysis Approach to Predicting Election Results: A Case Study of the 2015 Spanish Local Elections" de González-Bailón, S., Borge-Holthoefer, J., Moreno, Y. - publicată în PLOS ONE, 2016.

13. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment" de Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M. - publicată în International AAAI Conference on Weblogs and Social Media, 2010.
14. "Election Forecasting Using Twitter Sentiment Analysis: Towards a Credible Alternative" de Jungherr, A., Schoen, H., Jurgens, P. - publicată în Big Data & Society, 2016.
15. "Analyzing Twitter as a Political Discourse: A Sentiment Analysis" de Pak, A., Paroubek, P. - publicată în Proceedings of the Seventh International Conference on Language Resources and Evaluation, 2010.